UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"
PhD SCHOOL

PhD program in Statistics
Cycle: 31st
Disciplinary Field: SECS-S/01

Bayesian methods for the design and analysis of complex follow-up studies

Advisor: Pietro Muliere
Co-advisor: Lorenzo Trippa

PhD Thesis by
Andrea ARFÈ
ID number: 3000445

Academic year 2019/2020

To Sarah and my family
With special thanks to Pietro Muliere and Lorenzo Trippa

"We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction."
Otto Neurath

Abstract

My doctoral research focused on two specific topics: i) models for the analysis of multi-state time-to-event data; and ii) decision-theoretic approaches for the design of clinical trials with a survival endpoint. For the first, I developed stochastic processes useful for the Bayesian non-parametric analysis of follow-up studies where patients may experience multiple events relevant to their prognosis. For the second, I developed an approach that uses data from early clinical trials to specify the statistical test used in a confirmatory survival study, accounting for the possible failure of standard assumptions. In this thesis, I describe 3 research papers that report my contributions. Part of my work has been conducted while a visiting researcher at the Dana-Farber Cancer Institute, Boston, Massachusetts (United States of America).

Contents

Chapter 1

Introduction

My research interests focus on the development of Bayesian methods for biomedical studies. Both the design and analysis phase of such studies are fundamental for their success (Fisher et al., 1997; Cox and Donnelly, 2011). The Bayesian paradigm provides a great practical advantage in both, thanks to its intrinsic ability to combine information from multiple sources (Ashby and Smith, 2000; Dunson, 2001; Spiegelhalter et al., 2004). In addition, powerful computational tools permit to estimate and check complicated Bayesian models in many applications (Gelman et al., 2013).

In my doctoral research, I developed novel approaches for the design and analysis of complex survival studies. Specifically, I proposed new Bayesian non-parametric models for the analysis of follow-up studies whose participants may experience multiple events relevant to their prognosis (**Topic 1**). I also proposed a new framework - based on Bayesian decision-theory - for the design of randomized clinical trials with a survival end-point. (**Topic 2**). My contributions are described in **3 papers - 1 accepted for publication, 1 with invited revisions, and 1 under review**.

For **Topic 1**, I developed stochastic processes useful for the Bayesian non-parametric analysis of follow-up studies where i) there may be multiple competing endpoints (**Paper 1**) or ii) patients may progress through several disease states (**Paper 2**). Compared to parametric models, non-parametric models are less tied to restrictive assumptions that may give a false sense of posterior certainty (Hjort et al., 2010; Phadia, 2013; Ghosal and van der Vaart, 2017).

Instead, for **Topic 2** I developed a decision-theoretic procedure to specify the statistical test to be used the final analysis of a confirmatory survival study. The

procedure accounts for the possible failure of standard assumptions by leveraging data from past early-stage trials (**Paper 3**). Here, decision theory provides a coherent framework to both i) incorporate prior data in the design of a new experiment (Lindley, 1997; DeGroot, 2005; Parmigiani and Inoue, 2009) and ii) satisfy the requirements of pharmaceutical regulatory agencies (Ventz and Trippa, 2015).

The reminder of the thesis is structured as follows. In **Section 2** I provide a preliminary overview of the theoretical concepts used in **Papers 1-3**. In **Section 3**, I summarize the motivation and significance of the 3 research manuscripts, highlighting my contributions to each. In **Sections 4-6** I report the full text of **Papers 1-3**, respectively. Finally, in **Section 7** I provide some concluding remarks and describe future research. In the **Appendix**, I summarize **3 additional manuscripts (1 accepted for publication, 2 under review)** related to other applied research projects (which thus I do not report here in full). I contributed to these works while a visiting researcher at the Data Science Department, Dana-Farber Cancer Institute, Boston, Massachusetts (U.S.A.).

Chapter 2

Bayesian nonparametrics, semi-Markov processes, and decision theory for experimental designs: a brief overview

2.1 The Bayesian non-parametric approach

Contrary to the classic parametric case, which only deals with finite dimensional parameters (Berger, 2013; Gelman et al., 2013), the Bayesian non-parametric approach relies on prior probability distributions with an infinite dimensional support (Ferguson, 1973; Müller and Mitra, 2013; Ghosal and van der Vaart, 2017).

From an abstract point of view, a Bayesian non-parametric prior distribution is a probability measure on $M(\mathcal{X})$, the space of all probability measures on the sample space $\mathcal{X}$ (typically a metrizable Polish space) (Ghosal and van der Vaart, 2017).

Historically, the construction of one such prior was regarded as mathematically intractable, as it requires the definition of a *random probability measure*. This is a stochastic process with index set $\mathcal{B}$, the Borel σ-algebra of $\mathcal{X}$, and sample paths that form elements of $M(\mathcal{X})$.

Ferguson's seminal 1973 paper (Ferguson, 1973) represented a turning point for this research area. There, Ferguson introduced the *Dirichlet process*, a random probability measure that has since become widely used in Bayesian applications

(Ghosal and van der Vaart, 2017).

Definition 2.1.1 (Ferguson (1973)). *Fix $P_0 \in \mathcal{M}(\mathcal{X})$ and $k > 0$. A stochastic process $(P(A) : A \in \mathcal{B})$ with values in $[0,1]$ is called a* Dirichlet process *$Dir(k, P_0)$ if, for all measurable partitions $A_1, \ldots, A_n$ of $\mathcal{X}$, it is*

$$(P(A_1), \ldots, P(A_n)) \sim Dirichlet\,(kP_0(A_1), \ldots, kP_0(A_n))\,.$$

In his paper, Ferguson (1973) shows that the Dirichlet process is well-defined by appealing to conditions related to Kolmogorov's Extention Theorem (Çinlar, 2011, Theorem 4.18). He also shows that i) $\mathbb{E}[P(A)] = P_0(A)$ for all $A \in \mathcal{B}$, and ii) $\mathrm{Var}(P(A))$ is a decreasing function of k. He also shows that the distribution $\mathrm{Dir}(k, P_0)$ is *conjugate*: if $P \sim \mathrm{Dir}(k, P_0)$ and $X_1, \ldots, X_n$ is an independent and identically distributed sample from P, then the conditional distribution of P given $X_1, \ldots, X_n$ is $P \sim \mathrm{Dir}(k+n, P_0 + \sum_{i=1}^n \delta_{X_i})$ (where δ_x is the point-mass measure at x).

Building on Ferguson's work, Doksum (1974) introduced the following general definition of a random probability measure.

Definition 2.1.2 (Doksum (1974)). *A stochastic process $(P(A) : A \in \mathcal{B})$ is a* random finitely additive probability measure *if*

1. *$P(A) \in [0,1]$ almost surely for all $A \in \mathcal{B}$;*
2. *$P(\mathcal{X}) = 1$ almost surely;*
3. *for all $n \geq 1$ and all collections of disjointed sets $(A_{i,j} : 1 \leq j \leq m_i)$, $i = 1, \ldots, n$, the random vector*

$$\left(P\left(\cup_{j=1}^{m_1} A_{1,j}\right), \ldots, P\left(\cup_{j=1}^{m_n} A_{n,j}\right)\right)$$

has the same distribution as

$$\left(\sum_{j=1}^{m_1} P(A_{1,j}), \ldots, \sum_{j=1}^{m_n} P(A_{n,j})\right).$$

If P also satisfies the following condition, then it is called a random (σ-additive) probability measure*:*

4. if $A_k \in \mathcal{B}$ for all $k \geq 1$ and $A_k \downarrow \emptyset$, then $P(A_k) \to 0$ in distribution.

If $\mathcal{X} = (a, b) \subseteq \mathbb{R}$ is an interval ($-\infty \leq a < b \leq +\infty$), a random probability measure can be equally characterized by a *random distribution function*, another concept introduced by Doksum (1974).

Definition 2.1.3 (Doksum (1974))**.** *A stochastic process $(F(t) : t \in (a, b))$ is a* random distribution function *if, with probability 1,*

1. the function $t \mapsto F(t)$ is non-decreasing and right-continuous;

2. $\lim_{t \to a+} F(t) = 0$ and $\lim_{t \to b-} F(t) = 1$.

Since the contributions of Ferguson (1973) and Doksum (1974), the number of such processes introduced in the literature has grown substantially. For an review that covers diverse areas of application, see Hjort et al. (2010); Müller and Mitra (2013); Phadia (2013); Müller et al. (2015); Mitra and Müller (2015), and Ghosal and van der Vaart (2017).

2.2 Neutral-to-the-right processes

Here, I will focus on random probability measures defined on the sample space $\mathcal{X} = (0, +\infty)$. This are fundamental in the Bayesian non-parametric analysis of *time-to-event* or *survival data* (Ghosal and van der Vaart, 2017, Chapter 13). Here, observations are the (possibly censored) times elapsed before the onset of an event, such as the death of a patient in a clinical trial (Kalbfleisch and Prentice, 2002; Aalen et al., 2008), or the breakdown of a machine in an engineering study (Singpurwalla, 2006).

More specifically, I will focus on *Neutral-to-the-right processes*, a broad family of random distribution functions first introduced by Doksum (1974).

Definition 2.2.1 (Doksum (1974), Definition 3.1)**.** *The random distribution function $F(t)$ $(t > 0)$ is* neutral-to-the-right *if for all $n \geq 1$ and real numbers $0 < t_1 < t_2 < \cdots < t_n < +\infty$ the quantities*

$$F(t_1), \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, \ldots, \frac{F(t_n) - F(t_{n-1})}{1 - F(t_{n-1})}$$

are independent (with the convention that $0/0 = 0$).

Neutral-to-the-right distributions can be characterized by means of stochastic processes with independent increments, also know as *additive* (Sato, 1999) or *Lévy processes* (Doksum, 1974):

Definition 2.2.2. *A real-valued stochastic process* $(Z(t) : t \in (0, +\infty))$ *is called a* Lévy process *if, with probability 1,*

1. *the function* $t \mapsto Z(t)$ *is non-decreasing and right-continuous;*
2. $Z(0) = 0$ *and* $\lim_{t \to +\infty} Z(t) = +\infty$*;*
3. $Z(t)$ *has independent increments, i.e. for all* $0 < t_1 < \cdots < t_n$*, the quantities* $Z(t_1), Z(t_2) - Z(t_1), \ldots, Z(t_n) - Z(t_{n-1})$ *are independent.*

Intuitively, a distribution $F(t)$ is neutral-to-the-right if it has independent normalized increments. Indeed, Doksum (1974) showed that neutral-to-the-right processes are transformations of Lévy processes, which have independent increments.

Theorem 2.2.1 (Doksum (1974), Theorem 3.1)**.** *A random distribution function* $F(t)$ *is neutral to the right if and only if* $F(t) = 1 - \exp(-Z(t))$*, where* $(Z(t) : t \in \mathbb{R})$ *is a* Lévy process.

Any Lévy process $(Z(t) : t \in \mathbb{R})$ can be decomposed in a deterministic component and two independent jump process (Sato, 1999):

$$\begin{aligned} Z(t) &= d(t) + Z_f(t) + Z_r(t), \\ Z_f(t) &= \sum_{t_j \in \mathcal{J}} Z_{f,j} I\left\{t_j \leq t\right\}, \\ Z_r(t) &= \sum_{j} Z_{r,j} I\left\{T_j \leq t\right\}, \end{aligned}$$

where: $d(t)$ is a deterministic non-decreasing, right-continuous function; Z_f has jumps only at a countable number of fixed discontinuities $\mathcal{J} = \{t_1, t_2, \ldots\}$; the size of the jump at t_j is random and equal to $Z_{r,j}$; instead, $Z_r(t)$ has a countable number of jumps, which occur at random locations T_j and have random size $Z_{r,j}$. All the $Z_{f,j}$, $Z_{r,j}$, and T_j are independent.

The distribution of $Z_r(t)$ is characterized by the *log-Laplace transform*

$$\log \mathbb{E}\left[\exp(-\lambda Z_r(t))\right] = -\int_0^{+\infty} \left[1 - \exp(-\lambda s)\right] \nu_t(ds) \tag{2.1}$$

for all $t \in \mathbb{R}$ and $\lambda > 0$, where ν_t is the *Lévy measure* of $Z_r(t)$. This is a measure on $(0, +\infty)$ such that

$$\int_0^{+\infty} \min(s, 1)\nu_t(ds) < +\infty$$

for all $t > 0$ (Sato, 1999).

As a consequence, to specify a neutral-to-the-right prior on a distribution function $F(t)$, it is sufficient to specify i) the deterministic component $d(t)$, ii) the position t_j of each fixed jump of $Z_f(t)$, together with the distribution of its size $Z_{f,j}$; and iii) the Lévy measure ν_t of $Z_r(t)$.

Example 2.2.1 (Dirichlet process prior)**.** *Let P be a probability measure on $(0, +\infty)$ with distribution function $F(t)$. Then Dirichlet prior distribution $P \sim Dir(k, P_0)$ can be obtained as follows (Walker and Muliere, 1997, Remark 1; Ghosal and van der Vaart, 2017, Example 13.11). First, $F(t)$ is assumed a neutral-to-the-right process with $d(t) \equiv 0$. Second, if $\mathcal{J} = \{t_1, t_2, \dots, \}$ are the atoms of P_0, $0 < t_1 < t_2 < \dots$, the distribution of the jump $Z_{f,j} = F(t_j) - F(t_j-)$ is determined by*

$$1 - \exp(-Z_{f,j}) \sim Beta\left(kP_0(\{t_j\}), kP_0((t_j, +\infty))\right).$$

Third, the Lévy measure of $Z_c(t)$ is

$$d\nu_t(v) = dv \frac{k}{1 - e^{-v}} \int_0^t \exp\left(-vkP_0((s, +\infty))\right) dP_{0,c}(s),$$

where $P_{0,c}$ is the non-atomic component of P_0.

The moments $\mathbb{E}[F(t)^k]$, $k \geq 1$, all exist finite and can be obtained from the distributions of the $Z_{f,j}$ and from the Lévy measure ν_t. For example, if $Z(t)$ has no fixed jumps, then

$$\mathbb{E}[F(t)] = 1 - \exp\left(-\int_0^{+\infty} \left[1 - \exp(-s)\right] \nu_t(ds)\right).$$

For neutral-to-the-right processes without fixed jumps, Epifani et al. (2003) also show how to compute the moments of a functional $I = \int_0^{+\infty} g(t)dF(t)$, where $g(t)$ is a increasing function of t, from the Lévy measure ν_t.

The relevance of neutral-to-the-right processes in the Bayesian non-parametric framework is due to the results of Ferguson (1974) and Ferguson and Phadia (1979).

They establish that neutral-to-the-right prior processes are conjugate: if $F(t)$ is neutral-to-the-right and $X_1, \ldots, X_n$ is an independent and identically distributed sample from $F(t)$, then $F(t)$ is neutral-to-the-right also conditional on $X_1, \ldots, X_n$. In addition, this is true also when some observations are (right) censored: if $F(t)$ is neutral-to-the-right and X has distribution $F(t)$, then $F(t)$ is neutral-to-the-right conditional on $X > x$ for all fixed $x > 0$.

The conjugacy of neutral-to-the-right prior processes makes them very appealing for the analysis of survival data (Ghosal and van der Vaart, 2017, Chapter 13). In practice, neutral-to-the-right processes can be simulated by means of Monte Carlo procedures such as those of Damien et al. (1995), Walker and Damien (1998a), and Walker and Damien (1998b). See also Lee (2007).

From the predictive perspective, neutral-to-the-right processes can be fully characterized through an extension of *Johnson's sufficientness postulate* (Zabell et al., 1982). Johnson discovered that the Dirichlet distribution and process are fully characterized by a set of predictive distributions on discrete cells of the form $P(X_{n+1} = k|X_1, \ldots, X_n) = f_k(n_k)$, where n_k is the number of observations $X_1, \ldots, X_n$ in the k-th cell. Walker and Muliere (1999) show neutral-to-the-right processes are characterized by an extension of Johnson's construction that distinguishes between exact and censored. For a different characterization, see Muliere and Walker (2000).

To conclude this section, note that any Lévy processe $Z(t)$ on $\mathcal{X} = (0, +\infty)$ can be interpreted as the cumulative distribution function of a Completely Random Measure (c.f. Kingman, 1967; Ghosal and van der Vaart, 2017, Appendix J1) on the same sample space. From this perspective, James (2002, 2006) generalizes the notion of neutral-to-the-right processes to arbitrary Polish sample spaces.

2.3 The beta-Stacy process prior

The *beta-Stacy process* of Walker and Muliere (1997) is an important neutral-to-the-right process prior widely used for the analysis of time-to-event data. Some applications and generalizations include the works of Amerio et al. (2004); Muliere et al. (2005); Bulla and Muliere (2007); Bulla et al. (2009); Rigat and Muliere (2012), and Peluso et al. (2017).

Definition 2.3.1 (beta-Stacy process prior)**.** *Let $c(t) > 0$ for all $t > 0$ and let*

$F_0(t)$ a fixed distribution on $(0, +\infty)$. A neutral-to-the-right process $F(t) = 1 - \exp(-Z(t))$, $Z(t) = d(t) + Z_f(t) + Z_c(t)$, is a beta-Stacy process $BS(c, F_0)$ *if: i) the deterministic part $d(t)$ is null, i.e. $d(t) \equiv 0$; ii) if $\mathcal{J} = \{t_1, t_2, \ldots, \}$ are the atoms of F_0, $0 < t_1 < t_2 < \ldots$, the distribution of the jump $Z_{f,j} = F(t_j) - F(t_j-)$ is determined by*

$$1 - \exp(-Z_{f,j}) \sim Beta\left(c(t_j)(F_0(t_j) - F_0(t_j-)), c(t_j)(1 - F_0(t_j))\right);$$

and iii) the Lévy measure of $Z_c(t)$ is

$$d\nu_t(v) = dv \frac{1}{1 - e^{-v}} \int_0^t c(s) \exp\left(-vc(s)(1 - F_0(s))\right) dF_{0,c}(s),$$

where $F_{0,c}$ is the continuous component of F_0.

Note that if F_0 is purely atomic, then $F(t)$ is a beta-Stacy process if and only if

$$F(t) = 1 - \prod_{t_j \leq t} \exp(-Z_{f,j}) = 1 - \prod_{t_j \leq t} (1 - U_j),$$

where $U_1, U_2, \ldots$ are independent and such that

$$U_j \sim \text{Beta}\left(c(t_j)(F_0(t_j) - F_0(t_j-)), c(t_j)(1 - F_0(t_j))\right).$$

In this case, Walker and Muliere (1997) call $F(t)$ a *discrete-time beta-Stacy process*. For every $k \geq 1$, the joint distribution of the jumps $F(t_j) - F(t_j-)$, $j = 1, \ldots, k$, is a *Generalized Dirichlet distribution* of Connor and Mosimann (1969), also called the *beta-Stacy distribution* by Mihram and Hultquist (1967).

A consequence of Definition 2.3.1 is that, if $F(t) \sim BS(c, F_0)$, then, infinitesimally,

$$\frac{dF(t)}{1 - F(t-)} \sim \text{Beta}(c(t)dF_0(t), c(t)(1 - F_0(t))). \tag{2.2}$$

Informally, this implies that $\mathbb{E}[dF(t)/(1 - F(t-))] = dF_0(t)/(1 - F_0(t-))$, so

$$\mathbb{E}[F(t)] = 1 - \prod_{s \in (0,t]} \left(1 - \mathbb{E}\left[\frac{dF(s)}{1 - F(s-)}\right]\right) = F_0(t),$$

where $\prod_{s \in (0,t]}$ is the *product integral* operator (Gill et al., 1990). Additionally, $\text{Var}(dF(t)/(1 - F(t-))$ is a decreasing function of $c(t)$ such that $\text{Var}(dF(t)/(1 - F(t-))) \to 0$ as $c(t) \to +\infty$. Hence, $c(t)$ controls the variance of $F(t)/(1 - F(t-))$

at time t, i.e. the variability of $F(t)/(1-F(t-))$ around its mean $F_0(t)/(1-F_0(t-))$ (Walker and Muliere, 1997).

Equation 2.2 also provides a link between the beta-Stacy process and the *beta process* of (Hjort, 1990). Indeed, if $F(t)$ is a beta-Stacy process, then its *cumulative hazard function*

$$H(t) = \int_{(0,t]} \frac{dF(t)}{1 - F(t-)}$$

is a beta process. The converse is also true; see Walker and Muliere (1997, Remark 2), Ghosal and van der Vaart (2017, Chapter 13).

The beta-Stacy process generalizes the Dirichlet process, as can be seen by comparing Definition 2.3.1 with Example 2.2.1. Specifically, assume that i) $P \sim Dir(k, P_0)$, where P_0 is a probability measure on $(0, +\infty)$; and ii) $F_0(t) = P_0((0,t])$, $F(t) = P((0,t])$ for all $t > 0$. Then $F \sim BS(c, F_0)$, where $c(t) = k$ for all $t > 0$ (Walker and Muliere, 1997).

From the results of Ferguson (1974) and (Ferguson and Phadia, 1979) on neutral-to-the-right processes, Walker and Muliere (1997) characterize the posterior distribution of $F \sim BS(c, F_0)$ conditional on a sample of possibly right-censored observations:

Theorem 2.3.1 (Theorem 4, Walker and Muliere 1997)**.** *Suppose $F \sim BS(c, F_0)$ and, given F, let X be an observation from F. Then, conditional on $X = x$ (exact observation) or $X > x$ (right-censored observation), it is $F \sim BS(c^*, F_0^*)$, where:*

1. *if $X > x$,*

$$F_0^*(t) = 1 - \prod_{s \in (0,t]} \left(1 - \frac{c(s)dF_0(s)}{c(s)dF_0(s) + I\{x \geq s\}}\right)$$
$$c^*(t) = \frac{c(t)(1 - F_0(t-)) + I\{x \geq s\}}{1 - F_0^*(t)};$$

2. *if $X = x$,*

$$F_0^*(t) = 1 - \prod_{s \in (0,t]} \left(1 - \frac{c(s)dF_0(s) + I\{x = s\}}{c(s)dF_0(s) + I\{x \geq s\}}\right) \tag{2.3}$$
$$c^*(t) = \frac{c(t)(1 - F_0(t-)) + I\{x > s\}}{1 - F_0^*(t)}; \tag{2.4}$$

More generally, the posterior distribution of F conditional on an independent sample of exact or right-censored observations can be obtained by repeated application of (i) or (ii).

A consequence of Theorem 2.3.1 is that the Bayes estimate of $F(t)$ under a quadratic loss functions, $\widehat{F}(t) = \mathbb{E}[F(t)|\text{data}]$, converges to the classical Kaplan-Meier estimate (Kalbfleisch and Prentice, 2002, Section 1.4) as $c(t) \to 0$; see Walker and Muliere (1997, Section 4.3).

From Theorem 2.3.1, it is also possible to compute the predictive distribution of a new set of uncensored observations $X_{n+1}, \ldots, X_{n+k}$ given past data $X_1, \ldots, X_n$ (which may be censored). Specifically, let $F^*_{0,h}(\cdot|X_{n+1}, \ldots, X_{n+h})$ be the posterior mean of F given $X_1, \ldots, X_{n+h}$ for all $h = 0, \ldots, k$ (for simplicity $X_1, \ldots, X_n$ are suppressed from the notation). These functions can be computed similarly as F^* in Equation 2.3. The following proposition shows that they determine the predictive distribution of $X_{n+1}, \ldots, X_{n+k}$.

Proposition 2.3.1. The law of $X_{n+1}, \ldots, X_{n+k}$ given $X_1, \ldots, X_n$ is the probability measure $\prod_{h=1}^{k} F^*_{0,h-1}(dx_h|x_{n+1}, \ldots, x_{n+h})$.

Proof. Proceeding by induction, suppose first that $k = 1$. Then the law of X_{n+1} given $X_1, \ldots, X_n$ is $F^*_{0,0}$, since for any any $y > 0$ it is

$$\begin{aligned}\mathbb{P}(X_{n+1} \leq y|X_1, \ldots, X_n) &= \mathbb{E}[\mathbb{P}(X_{n+1} \leq y|X_1, \ldots, X_n, F)|X_1, \ldots, X_n] \\ &= \mathbb{E}[F(y)|X_1, \ldots, X_n] \\ &= F^*_{0,0}(y),\end{aligned}$$

where the last equality follows from Theorem 2.3.1.

Suppose now that the thesis is true for $k \geq 1$. With a similar argument as above, it is seen that the law of X_{n+k+1} given $X_1, \ldots, X_{n+k}$ is $F^*_{0,k}(dx_{n+k+1}|X_{n+1}, \ldots, X_{n+k})$. By the inductive hypothesis, the law of $X_{n+1}, \ldots, X_{n+k}$ given $X_1, \ldots, X_n$ is

$$\prod_{h=1}^{k} F^*_{0,h-1}(dx_h|x_{n+1}, \ldots, x_{n+h}).$$

Thus, the law of $X_{n+1}, \ldots, X_{n+k}, X_{n+k}$ given $X_1, \ldots, X_n$ is

$$F^*_{0,k}(dx_{n+k+1}|x_{n+1}, \ldots, x_{n+k}) \cdot \prod_{h=1}^{k} F^*_{0,h-1}(dx_h|x_{n+1}, \ldots, x_{n+h}),$$

as needed. □

2.4 Predictive constructions based on reinforced urn models

The characterization of models and priors through a predictive approach is a fundamental, long studied problem in Bayesian statistics. Indeed, as often underlined by de Finetti, one can only express a subjective probability on observable facts, and probabilistic models are just a way to link past past experience with potential future observations. For a discussion, see Cifarelli and Regazzini (1982), Geisser (1993), Wechsler (1993), Bernardo and Smith (2000, Chapter 4), and Kallenberg (2010).

These fundamental problems have often been thought as mainly theoretical, with little practical applications. However, in the recent years they have received a renewed interested motivated by applications in Bayesian non-parametrics (Fortini and Petrone, 2012). Muliere et al. (2003) note how predictive constructions may even allow to implement Bayesian non-parametric inference without explicit knowledge of the prior.

From this perspective, many non-parametric prior processes available in the literature have been characterized from using *urn models* (Mahmoud, 2008). This are the prototypical example of stochastic processes with *reinforcement*, i.e. of processes where occurrence of an event increases the probability of future similar events (Coppersmith and Diaconis, 1986; Pemantle, 2007).

A seminal paper in this area of research is due to Blackwell and MacQueen (1973), who characterized the Dirichlet process $Dir(k, P_0)$ using a *(generalized) Pólya urn*. This is an urn that contains k coloured balls, where each $x \in \mathcal{X}$ is interpreted as a different "color". For all (measurable) $A \subseteq \mathcal{X}$, $P_0(A)$ is the proportion of balls in the urn of color $x \in A$.

The urn evolves in the following steps, which can be repeated indefinitely: a ball is randomly extracted, its color is noted, and then it is replaced in the urn; the extracted color is then reinforced by introducing another ball of the same color in the urn.

As an example, at the first draw the probability of obtaining a color in $A \subseteq \mathcal{X}$ is $P_0(A)$. If the color x is extracted, the number of balls in the urn becomes $k+1$, while the number of balls of color in A becomes $kP_0(A) + \delta_x(A)$. At the second draw, the probability of observing a color in A is $(kP_0(A) + \delta_x(A))/(k+1)$. If the color y is extracted, this probability becomes $(kP_0(A) + \delta_x(A) + \delta_y(A))/(k+2)$, and

so on.

Definition 2.4.1 (Blackwell and MacQueen (1973)). *If X_n is the n-th color extracted from the urn as described above, then the sequence $(X_n : n \geq 1)$ is called a* Pólya sequence *with parameters k and P_0.*

The main result of Blackwell and MacQueen (1973) is the following:

Theorem 2.4.1 (Blackwell and MacQueen (1973)). *If $(X_n : n \geq 1)$ is a Pólya sequence with parameters k and P_0, then it is exchangeable and its associated de Finetti measure is the law of a $Dir(k, P_0)$ process. In other words, there exists a random probability $Q \sim Dir(k, P_0)$ on $\mathcal{X}$ such that, conditional on Q, the X_n are independent, all with distribution Q.*

Following the contribution of Blackwell and MacQueen (1973), many other nonparametric prior processes were characterized using urn models, including Pólya trees (Mauldin et al., 1992), mixtures of Dirichlet processes (Fortini et al., 2016), the Enriched Dirichlet Process (Wade et al., 2011), the Pitman-Yor process Pitman (1996), and the time-varying Pitman-Yor process Caron et al. (2017). In particular, (Muliere et al., 2000) provide a construction of general neutral-to-the-right processes in discrete time.

The characterization of the discrete-time beta-Stacy process provided by (Muliere et al., 2000), which extends a previous construction by (Walker and Muliere, 1997), will play a special role in the sequel. Briefly, suppose that F_0 is a purely-atomic distribution on $(0, +\infty)$ with atoms $0 < t_1 < t_2 < \cdots < t_j < \cdots$. Also suppose that $c(t_j) > 0$ for all $j = 1, 2, \ldots$. Associate every t_j with a Pólya urn V_j that contains $c(t_j)(F_0(t_j) - F_0(t_j-))$ black balls and $c(t_j)(1 - F_0(t_j))$ white balls. Starting from V_1, for $k \geq 1$ sample a ball from V_k. If its color is white, continue sampling from V_{k+1}, otherwise set $X_1 = t_k$ and return to V_1 after reinforcing all visited urns. Restarting from V_1, the process is repeated to generate X_2, X_3, X_4, and so on (see Figure 2.1 for an illustration). Under these conditions, the results of Muliere et al. (2000) imply that i) the sequence $(X_n : n \geq 1)$ is exchangeable and ii) its de Finetti measure is the $BS(c, F_0)$ distribution.

Figure 2.1: Illustration of the urn process characterizing the discrete-time beta-Stacy distribution. At the start of the process (line 1), the first black ball is extracted from the urn V_2, generating the first observation $X_1 = t_2$. After reinforcing the all urns (line 2), the process is repeated. The first black ball is extracted from urn V_4, generating the observation $X_2 = t_4$. The process the repeats again (line 3), generating the observation $X_3 = t_3$. Repeating this an infinite number of times generates an exchangeable sequence $(X_n : n \geq 1)$ with beta-Stacy de Finetti measure.

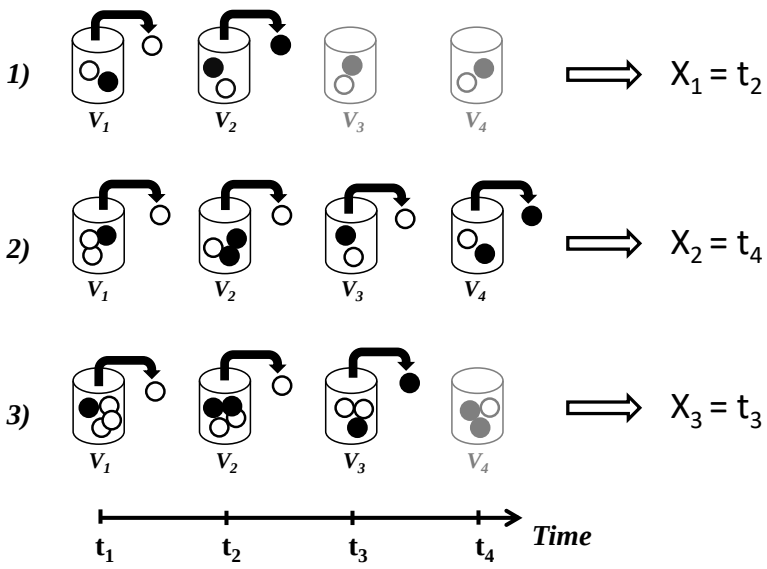
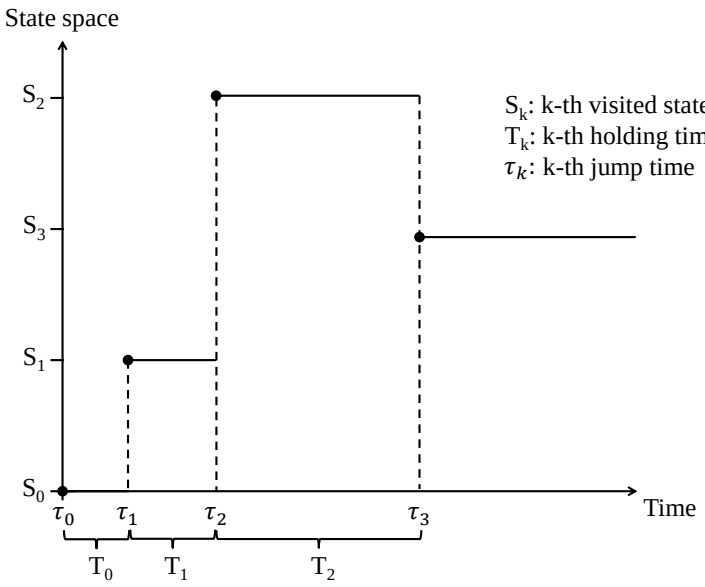


Figure 2.2: Representation of the sample paths of a Semi-Markov process. The process changes value only at the jump times $0 < \tau_1 < \tau_2 < \ldots$.

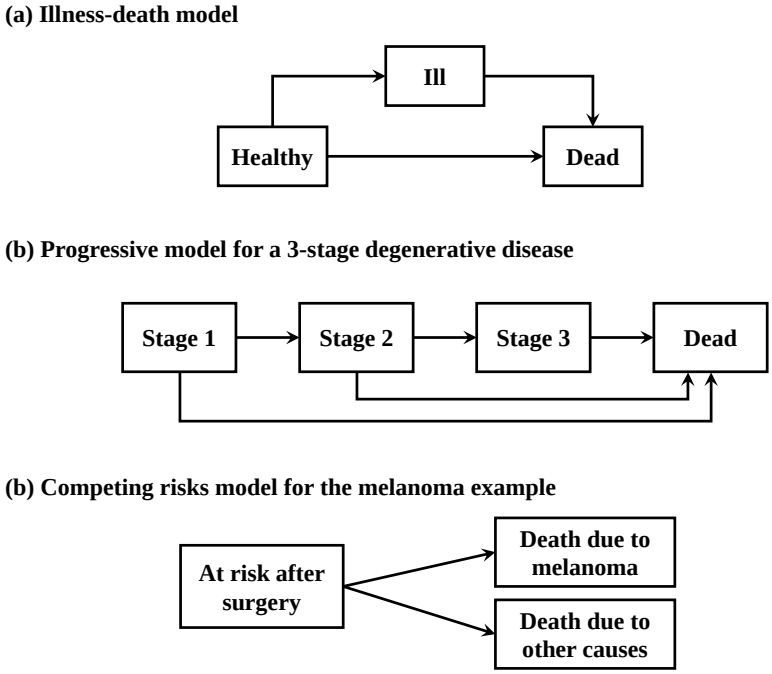


2.5 Semi-Markov processes and competing risks

A *(discrete state) semi-Markov process* is a stochastic process $(S_t)_{t\geq 0}$ with a countable state space $E = \{e_0, e_1, e_2, ...\}$ and piecewise-constant trajectories with jumps at random times $0 < \tau_1 < \tau_2 < \ldots$ (see Figure 2.2). Its defining property is that the sequence of values S_{τ_j} at the jump times must form a homogeneous Markov Chain. However, in contrast with a Markov Chain, the *holding times* $T_k = \tau_k - \tau_{k-1}$ (i.e. the times spent in each visited state) can have an arbitrary distribution (Çinlar, 1969).

Because of their flexibility, semi-Markov processes are widely used to predict the evolution of phenomena that progress through different discrete states. They have been applied in research areas as diverse as longitudinal and time-series (Bulla and Bulla, 2006), Finance and actuarial sciences (Janssen and Manca, 2007), Biology (Barbu et al., 2004), and reliability analysis (Mitchell et al., 2011). In applications, interest is typically on estimating the probability that the process performs specific transitions or the distribution of the holding times and predict its future evolution.

Figure 2.3: Examples of simple semi-Markov models from the biomedical setting: (a) illness-death model; (b) progressive model for a 3-stage degenerative disease; (c) competing risks model for melanoma patients after radical resection. Boxes represent the possible states of the process, arrows represent the possible state transitions.



In biomedical fields, semi-Markov models are used to model many multi-state disease processes (Andersen and Keiding, 2002; Meira-Machado et al., 2009). Here, two common semi-Markov models are the *illness-death model*, which is used to evaluate whether previously diseased patients have the same mortality risk as those who have been healthy (c.f. Figure 2.3, Panel a), and the *progressive disease model*, which is used to evaluate the prognosis of patients affected by a multi-stage disease, e.g. cancer patients, whose tumour may progress through 4 stages of increasing severity, or patients affected by a degenerative disease (c.f. Figure 2.3, Panel b).

In my work, I focused on another, widely used, semi-Markov model: the *competing risks model* (Kalbfleisch and Prentice, 2002; Andersen et al., 2002; Huber et al., 2004; Lau et al., 2009; Crowder, 2012). A *competing risk* is an event that

hinders the observation of another event of interest. For example, melanoma patients that undergo a radical resection of the tumour typically have a long survival. Consequently, in a follow-up study which aims to understand the impact of radical resection on survival, deaths due to melanoma may not be observed only because some other cause of mortality (e.g. death due to cardiovascular disease) has occurred first. To understand the impact of radical resection, it thus becomes necessary to account death due to melanoma and death due to other causes as competing risks (c.f. Figure 2.3, Panel c) (Shen et al., 2016).

2.6 Constrained decision problems for experimental designs

Consider the problem of planning a confirmatory clinical trial whose goal is to test whether an experimental drug is superior to a standard therapy. The design d of the experiment includes a description of all the analyses that will be performed in the study to generate a final recommendation.

To solve this problem, in the Bayesian decision-theoretic approach one would roughly proceed as follows (Parmigiani and Inoue, 2009; Berger, 2013). First, one would specify a probability model $f(Y|\theta, d)$ for the data Y generated by the experiment, together with a prior distribution $\pi(\theta)$ for the parameter set θ. A *utility function* $u(Y, \theta, d)$ would then be specified to measure the preference assigned to each possible outcome of the study. Finally, one would choose the *optimal Bayesian (OB) design* d_{OB}, i.e. the design d that maximizes the *expected utility*

$$U(d) = \int \int u(Y, \theta, d) f(Y|\theta, d) \pi(\theta) dY d\theta.$$

Beyond this simple example, the Bayesian decision-theoretic approach has been used to develop designs for a variety of clinical experiments; see for example Berry and Ho (1988); Muliere and Petrone (1993); Stallard et al. (1999); Ding et al. (2008); Trippa et al. (2012); Cellamare et al. (2016), and Ventz et al. (2018). Berry and Kadane (1997) discuss the use of randomization in clinical trials from the perspective of decision theory. Carlin et al. (1998) and Müller et al. (2007) instead discuss computational methods to select optimal designs.

The decision-theoretic framework is very appealing for the design of clinical trials. This is because: i) it facilitates the use of past data to plan a new experiment (Spiegelhalter et al., 2004); ii) it satisfies the likelihood principle, which makes trial data interpretation unrelated to preplanned stopping rules and interim analyses (Cornfield, 1966; Berry et al., 2010); and iii) it allows to use utility functions to design experiments that explicitly take in consideration the need of researchers to reach a decision on the basis of the generated data (Lindley, 1994; Anscombe, 1963).

Despite the usefulness of Bayesian approaches, most studies are designed from a frequentist perspective (Chevret, 2012). Concepts such as the Type I and II error rates are required to be part of the design in order to be approved by pharmaceutical regulatory agencies, such as the E.U. European Medicines Agency, or the U.S. Food and Drug Administration (Phillips and Haudiquet, 2003; US Food and Drug Administration, 1998).

For example, consider an experimental design that is proposed by an Investigator, and a Regulator that either approves or rejects the design. The Regulator requires that the design d proposed by the Investigator must satisfy a set of operating characteristics, say V. For instance, in the confirmatory randomized trial example, the Regulator might require the control of the type I error at a suitable α level (e.g. 5%), and a power larger than some threshold $1-\beta$ (e.g. 80%). In this case, $V = [0, \alpha] \times [1-\beta, 1]$. The operating characteristics of d relevant to the Regulator, denoted by $oc(d)$ are thus type I error rate $\alpha(d)$ and power $1-\beta(d)$ of the design d, i.e. $oc(d) = (\alpha(d), 1-\beta(d))$. The Regulator only accepts study designs such that $oc(d) \in V$.

Bayesian designs do not require considerations of frequentist operating characteristics. Indeed, it may well be $oc(d_{OB}) \notin V$, so the optimal Bayesian design d_{OB} could be rejected by the Regulator. In practice, simulations are commonly used to estimate the operating characteristics of Bayesian designs. If necessary, the Investigator then adjusts tuning parameters to obtain a design that with a pre-specified type I error probability. This approach is discussed for instance by Lewis and Berry (1994), who use an iterative adjustment of the utility function to obtain designs with pre-specified frequentist properties.

The *constrained decision approach* to experimental design provides an alternative paradigm that combines aspects of the frequentist and Bayesian paradigms (Ventz and Trippa, 2015; Ventz et al., 2017). In this approach, a fixed utility function is

used to represent the preferences of the Investigator, which is used to quantify the utility of different study designs. The selected design must then both i) maximize the expected utility of the experiment and ii) satisfy a set of constraints defined by a set of frequentist characteristics.

Formally, let D be the set of all possible designs d. In the constrained decision approach, the Investigator selects *constrained optimal design* d_{CO}, which is determined by

$$d_{CO} = \arg\max\{U(d) : d \in D \text{ such that } oc(d) \in V\},$$

the solution to a constrained optimization problem. Indeed, the optimum d_{CO} has the highest expected utility $U(d)$ for the Investigator among the set of designs that the Regulator allows to implement. Moreover, if $oc(d_{BO}) \in V$, then $d_{CO} = d_{BO}$.

The constrained optimal design can be interpreted as the Bayesian optimal design with respect to a utility function which obeys the requirements of the Regulator. Specifically, assume for simplicity that $u(Y, \theta, d)$ is bounded from below by $u_{\min} > -\infty$. Consider the utility function

$$u^*(Y, \theta, d) = u_{\min} + (u(Y, \theta, d) - u_{\min})I\{oc(d) \in V\}.$$

With this utility function, any design $d \in D$ that violates the Regulator's constraints has minimal utility $u_{\min}$ for the Investigator, whereas $u^*(Y, \theta, d) = u(Y, \theta, d)$ if $oc(d) \in V$. Additionally, if

$$U^*(d) = \int\int u^*(Y, \theta, d) f(Y|\theta; d)\pi(\theta) dY\, d\theta,$$

it can be easily shown that

$$d_{CO} = \arg\max\{U^*(d) : d \in D\}.$$

Note that, in the this approach, the Investigator is not required to give up the utility function $u(Y, \theta, d)$, model $f(Y|\theta; d)$, or prior distribution $\pi(\theta)$. Contrary to other approaches to control the frequentist operating characteristics of Bayesian designs, the constrained decision approach allows the Investigator to make use of genuine representations of his or her preferences and prior knowledge (Ventz and Trippa, 2015).

Chapter 3

Summary of manuscripts

Here I report a summary of the 3 papers presented in the remainder of this thesis, along with a description of my original contributions.

Paper 1 (Topic 1): *Reinforced urns and the subdistribution beta-Stacy process prior for competing risks analysis. (With Pietro Muliere and Stefano Peluso; Scandinavian Journal of Statistics 2019; 46:706-734).*

In clinical prognostic research with a time-to-event outcome, the occurrence of one of several competing risks often precludes the occurrence of another event of interest (Kalbfleisch and Prentice, 2002, Chapter 8). In such cases it is typically of interest to assess i) the probability that one of the considered competing risks occurs within some time interval and ii) how this probability changes in association with predictors of interest (Wolbers et al., 2009; Fine, 1999; Putter et al., 2007). In contrast with the frequentist literature (Andersen et al., 2012), the Bayesian literature on competing risks is still sparse.

In **Paper 1** (Section 4), I introduce the *subdistribution beta-Stacy process*, a generalization of the beta-Stacy process prior (c.f. Section 2.3) useful for the analysis of competing risks data (c.f. Section 2.5). In particular, I i) characterize the process from a predictive perspective by means of an urn model with reinforcement (c.f. Section 2.4), ii) show that it is conjugate with respect to right-censored data, and iii) highlight its relations with other prior processes for competing risks data. Additionally, I consider the subdistribution beta-Stacy process prior in a regression

model for competing risks data that, contrary to most others available in the literature, is not based on the proportional hazards assumption. I provide code in the R statistical language to implement the proposed model.

Contributions. I defined the subdistribution beta-Stacy process and proved all related theoretical results (e.g. conjugacy and predictive characterization). I also specified the regression model for competing risks presented in the paper and conducted the applied analyses. I collaborated with Stefano Peluso to develop the R code used to implement the models introduced in the paper. I wrote all drafts of the manuscript and curated its revisions in collaboration with Stefano Peluso and Pietro Muliere.

Paper 2 (Topic 1): *The semi-Markov beta-Stacy process: a Bayesian non-parametric prior for semi-Markov processes. (With Pietro Muliere and Stefano Peluso. Submitted manuscript)*

Because of their flexibility, Discrete-time semi-Markov processes (a generalization of Markov chains; Çinlar, 1969) are widely used to predict many phenomena that evolve through a sequence of discrete states. Applications include time-series and longitudinal data analysis (Bulla and Bulla, 2006), survival analysis and reliability (Barbu et al., 2004; Mitchell et al., 2011), finance and actuarial sciences (Janssen and Manca, 2007), and biology (Barbu and Limnios, 2009). Despite their usefulness, and in contrast with their continuous-time counterparts (Phelan 1990; Bulla and Muliere 2007; Zhao and Hu 2013), the literature on inferential or predictive approaches for discrete-time semi-Markov process is sparse (Barbu and Limnios, 2009, Chapter 4).

Extending the work in **Paper 1**, in **Paper 2** (Section 5) I introduce the *semi-Markov beta-Stacy process*, a stochastic process useful for the Bayesian non-parametric analysis of semi-Markov processes (c.f. Section 2.5). I show that the semi-Markov beta-Stacy process is conjugate with respect to data generated by a semi-Markov process. In addition, I characterize the predictive distributions of the semi-Markov beta-Stacy process as a reinforced random walk on a system of urns (c.f. Section 2.4). I also explore two generalizations of the semi-Markov beta-Stacy process.

Contributions. I defined the semi-Markov beta-Stacy process and proved all re-

lated theoretical results (e.g. conjugacy). I also introduced and studied all urn-based characterizations presented in the paper. I developed the R code used to implement the simulation study presented in the manuscript. Finally, I wrote all drafts of the manuscript and curated its revisions in collaboration with Stefano Peluso and Pietro Muliere.

Paper 3 (Topic 2): *Bayesian optimality of testing procedures for survival data in the non-proportional hazards setting. (With Brian Alexander and Lorenzo Trippa. Submitted manuscript)*

Researchers often use data generated by exploratory clinical studies to specify the protocol of randomized confirmatory phase III trials (Lindley, 1997; Gómez et al., 2014; Lee and Wason, 2018; Brody, 2016). Still, in most cases prior information is not used to specify in the protocol, as mandated by pharmaceutical regulatory agencies (US Food and Drug Administration, 1998), which hypothesis testing procedure will be used in the final analyses to provide evidence of treatment effects. Most standard tests for treatment effects used in randomized clinical trials with survival outcomes are based on the proportional hazards assumption. This often fails in practice, impairing the study's power (Royston and Parmar, 2013). Data from early exploratory studies may provide evidence of non-proportional hazards which can guide the choice of alternative tests in the design of confirmatory trials.

In **Paper 3** (Section 6) I study a test to detect treatment effects in a late-stage trial which accounts for the deviations from proportional hazards suggested by early-stage data. I derive the test as the solution of a constrained decision problem (c.f. Section 2.6): conditional on early-stage data, the test maximizes the predicted finite-sample power among all tests that control the frequentist Type I error rate of the late-stage study at a prespecified α level (as required by regulatory agencies; US Food and Drug Administration, 1998). More precisely, the test maximizes the Bayesian predictive probability of correctly rejecting the null hypothesis at the end of the confirmatory trial. I provide R code to implement the proposed test.

Contributions. Brian Alexander introduced me and Lorenzo Trippa to the problem of delayed treatment effects in clinical trials of immuno-oncology treatments - the motivating application of this work. I defined the testing procedure introduced

in the paper and proved all related theoretical results, with an important contribution from Lorenzo Trippa in checking the logic of the proofs. I developed all R code used to implemented the proposed approach and performed all applied analyses in the paper. I wrote all drafts of the manuscript and curated its revisions in collaboration with Brian Alexander and Lorenzo Trippa.

Chapter 4

Reinforced urns and the subdistribution beta-Stacy process prior for competing risks analysis

With Stefano Peluso and Pietro Muliere.
Scandinavian Journal of Statistics 2019; 46:706-734.
ArXiv manuscript: `https://arxiv.org/abs/1811.12304`

4.1 Introduction

In the setting of clinical prognostic research with a time-to-event outcome, the occurrence of one of several competing risks may often preclude the occurrence of another event of interest (Kalbfleisch and Prentice, 2002, Chapter 8). In such cases it is typically of interest to assess i) the probability that one of the considered competing risks occurs within some time interval and ii) how this probability changes in association with predictors of interest (Wolbers et al., 2009; Fine, 1999; Putter et al., 2007). For example, in a study of melanoma patients who received radical surgery such as that of Drzewiecki et al. (1980), interest may be on the risk of melanoma-related mortality or melanoma-unrelated mortality and their potential predictors. Here, melanoma-related and melanoma-unrelated death act as competing risks, since onset of one necessarily precludes the onset of the other (Andersen et al., 2012, Chapter 1).

Competing risks data has received widespread attention in the frequentist literature. It suffices to recall the comprehensive textbooks of Kalbfleisch and Prentice (2002), Pintilie (2006), Aalen et al. (2008), Lawless (2011), Andersen et al. (2012) and Crowder (2012). Putter et al. (2007), Wolbers et al. (2009), and Andersen et al. (2002) provide an introductory overview of standard approaches for competing risks data. Classical approaches to prediction in presence of competing risks focus on the *subdistribution function*, also known as the *cumulative incidence function*, which represents the probability that a specific event occurs within a given time period. Kalbfleisch and Prentice (2002, Chapter 8) describe a frequentist nonparametric estimator for the subdistribution function, while Fine and Gray, in their pivotal 1999 paper, introduced a semiparametric proportional hazards model for the subdistribution function. Fine and Gray (1999) and Scheike et al. (2008) considered alternative semiparametric estimators, whilst Larson and Dinse (1985), Jeong and Fine (2007), and Hinchliffe and Lambert (2013) considered parametric regression models for the subdistribution function.

In contrast with the frequentist literature, the Bayesian literature on competing risks is still sparse, although several relevant contributions can be identified. Ge and Chen (2012) introduced a semiparametric model for competing risks by separately modelling the subdistribution function of the primary event of interest and the conditional time-to-event distributions of the other competing risks. They modelled the baseline subdistribution hazards and the cause-specific hazards by means of a gamma process prior (see Nieto-Barajas and Walker, 2002 and Kalbfleisch and Prentice, 2002, Section 11.8). De Blasi and Hjort (2007) suggested a semiparametric proportional hazards regression model with logistic relative risk function for cause-specific hazards. For inference, they assign the common baseline cumulative hazard a beta process prior (Hjort, 1990). With the same approach, Hjort's extension of the beta process for nonhomogeneous Markov Chains (Hjort, 1990, Section 5) may be considered as a prior distribution on the set of cause-specific baseline hazards in a more general multiplicative hazards model (see Andersen et al., 2012, Chapter III and Lawless, 2011, Chapter 9). In the beta process for nonhomogeneous Markov Chains the individual transition hazards are necessarily independent (Hjort, 1990, Section 5). The beta-Dirichlet process, a generalization of the beta process introduced by Kim and Gray (2012), relaxes this assumption by allowing for correlated hazards. Kim and Gray (2012) use the beta-Dirichlet process to define a semipara-

metric semi-proportional transition hazards regression model for nonhomogeneous Markov Chains which, in the competing risks setting, could be used to model the cause-specific hazards. With the same purpose, Chae et al. (2013) proposed a nonparametric regression model based on a mixture of beta-Dirichlet process priors.

In this paper we introduce a novel stochastic process, a generalization of Walker and Muliere's beta-Stacy process (Walker and Muliere, 1997), which represents a nonparametric prior distribution (i.e. a probability distribution on an infinite-dimensional space of distribution functions; see Ferguson (1973); Hjort et al. (2010); Müller and Mitra (2013)) useful for the Bayesian analysis of competing risks data. This new process, which we call the *subdistribution beta-Stacy* process, is conjugate with respect to right-censored observations, greatly simplifying the task of performing probabilistic predictions. We will also use the subdistribution beta-Stacy process to specify a Bayesian competing risks regression model useful for making prognostic predictions for individual patients. Contrary to most available regression approaches for competing risks, ours is not based on the proportional hazards assumption. As an illustration, we implement our model to analyse a classical dataset relating to survival of patients after surgery for malignant melanoma (Andersen et al., 2012, Chapter 1). Throughout the paper, our perspective is Bayesian nonparametric because: i) the Bayesian interpretation of probability is especially suited for representing uncertainty when making predictions (de Finetti, 1937; Singpurwalla, 1988); ii) Bayesian nonparametric models typically provide a more honest assessment of posterior uncertainty than parametric models, as the formers are less tied to potentially restrictive and/or arbitrary parametric assumptions which may give a false sense of posterior certainty (Müller and Mitra, 2013; Hjort et al., 2010; Phadia, 2013; Ghosal and van der Vaart, 2017).

To characterize the subdistribution beta-Stacy process we adhere to the *predictive approach*, a framework championed by de Finetti (1937) which is receiving renewed attention in statistics and machine learning as a useful tool for constructing Bayesian nonparametric priors (Fortini and Petrone, 2012; Orbanz and Roy, 2015). In the predictive approach, both the model and the prior are implicitly characterized by first specifying the predictive distribution of the observable quantities and then by appealing to results related to the celebrated de Finetti Representation Theorem (Walker and Muliere, 1999; Muliere et al., 2000; Epifani et al., 2002; Muliere et al., 2003; Bulla and Muliere, 2007; Fortini and Petrone, 2012). In our context,

the predictive distribution represents a specific rule prescribing how probabilistic predictions for a new patient should be performed after observing the experience of other similar (exchangeable) patients. This makes the predictive approach especially suited for our purposes: as our focus is on making prognostic predictions for individual patients, it seems natural to focus directly on the predictive distribution and its properties. Additionally, the predictive approach avoids some conceptual difficulties arising when specifying prior distributions for unobservable quantities (such as cause-specific hazards or other finite- or infinite-dimensional parameters). In fact, as often underlined by de Finetti and others, one can only express a subjective probability on observable facts; the role of unobservable quantities is just to provide a link between past experience and the probability of future observable facts (de Finetti, 1937; Singpurwalla, 1988; Wechsler, 1993; Cifarelli and Regazzini, 1996; Bernardo and Smith, 2000, Chapter 4; Fortini and Petrone, 2012).

The predictive rule underlying the subdistribution beta-Stacy process will be described in terms of the laws determining the evolution of a *reinforced urn process* (Muliere et al., 2000). Urn models have been used to characterize many common nonparametric prior processes. Classic examples include the use of a Pólya urn for generating a Dirichlet process (Blackwell and MacQueen, 1973), Pólya trees (Mauldin et al., 1992), and a generalised Pólya-urn scheme for sampling the beta-Stacy process (Walker and Muliere, 1997); Fortini and Petrone (2012) provide references to other modern examples. From this perspective, reinforced urn processes provide a general framework for building such urn-based characterizations. In fact, Muliere et al. (2000) and Muliere and Walker (2000) showed how reinforced urn processes can be used to characterize Pólya trees, the beta-Stacy process, and even general neutral-to-the-right processes (Doksum, 1974). Reinforced urn processes have also been applied for Bayesian nonparametric inference in many contexts, from survival analysis (Bulla et al., 2009) to credit risk (Peluso et al., 2015), thanks to their flexibility in modelling systems evolving through a sequence of discrete states.

The main idea behind reinforced urn processes is that of reinforced random walk, introduced by Coppersmith and Diaconis (1986) for modeling situations where a random walker has a tendency to revisit familiar territory; see also Diaconis (1988) and Pemantle (1988, 2007). In detail, a reinforced urn process is a stochastic process with countable state-space S. Each point $x \in S$ is associated with an urn containing coloured balls. The possible colors of the ball are represented by the elements of

the finite set E. Each urn $x \in S$ initially contains $n_x(c) \geq 0$ balls of color $c \in E$. The quantities $n_x(c)$ need not be integers, although thinking them as such simplifies the description of the process. For a fixed initial state x_0, recursively define the process as follows: i) if the current state is $x \in S$, then a ball is sampled from the corresponding urn and replaced together with a fixed amount $m > 0$ of additional balls of the same color; hence, the extracted color is "reinforced", i.e. made more likely to be extracted in future draws from the same urn (Coppersmith and Diaconis, 1986; Pemantle, 1988, 2007); ii) if $c \in E$ is the color of the sampled ball, then the next state of the process is $q(x, c)$, where $q : S \times E \to S$ is a known function, called the *law of motion* of the process, such that for every $x, y \in S$ there exists a unique $c(x, y) \in E$ satisfying $q(x, c(x, y)) = y$. For our purposes, the sequence of colors extracted from the urns will represent the history of a series of sequentially observed patients. The "reinforcement" of colors will then correspond to the notion of "learning from the past" that allows predictions to be performed and which is central in the Bayesian paradigm (Muliere et al., 2000, 2003; Bulla and Muliere, 2007; Peluso et al., 2015).

Before continuing, we must remark on the choice between continuous versus discrete time scales in the modelling of time-to-event distributions. In many, if not all, real applications, event times are not observed or available on a continuous time scale. Rather, they are either i) intrinsically discrete or ii) they are discrete because they arise from the coarsening of continuous data due to imprecise measurements (Kalbfleisch and Prentice, 2002, Chapter 2; Tutz and Schmid, 2016, Chapter 1; Allison, 1982; Guo and Lin, 1994). For this reason, throughout the paper we assume that the time axis has been pre-emptively discretized according to the fixed partition $(0, \tau_1]$, $(\tau_1, \tau_2]$, ..., $(\tau_{t-1}, \tau_t]$, ... (representing, say, successive days, months, years, etc.) implied by the measurement scale of event times in the considered application. Specifically, we assume that events can only occur at the times $\tau_1 < \tau_2 < \ldots$, in case (i), or that it is only possible to known in which intervals among $(0, \tau_1]$, $(\tau_1, \tau_2]$, ..., $(\tau_{t-1}, \tau_t]$, ... they occur, in case (ii). For notational simplicity, and without loss of generality, we also assume that any time-to-event variable $T > 0$ takes values in the set of positive integers $t \geq 1$: the observation that $T = t$ represents either the fact that the event occurred at time τ_t, in case (i), or during $(\tau_{t-1}, \tau_t]$, in case (ii).

4.2 The subdistribution beta-Stacy process

Suppose that the positive discrete random variable $T \in \{1, 2, \ldots\}$ represents the time until an at-risk individual experiences some event of interest (e.g. time from surgery for melanoma to death). If the distribution of T is unknown, then, in the Bayesian framework, it may be assigned a nonparametric prior to perform inference. In other words, it may be assumed that, conditionally on some random distribution function G defined on $\{0, 1, 2, \ldots\}$, T is distributed according to G itself: $P(T \leq t|G) = G(t)$ for all $t \geq 0$, or also $P(T = t|G) = \Delta G(t)$, where $\Delta G(0) = G(0) = 0$ and $\Delta G(t) = G(t) - G(t-1)$ for all integers $t \geq 1$. Thus the random distribution function G assumes the role of an infinite-dimensional parameter, while its distribution corresponds to the nonparametric prior distribution. The *beta-Stacy process* of Walker and Muliere (1997) is one of such nonparametric priors which has received frequent use. Specifically, a random distribution function G on $\{0, 1, 2, \ldots\}$ is a discrete-time beta-Stacy process with parameters $\{(\beta_t, \gamma_t) : t \geq 1\}$, where

$$\lim_{t \to +\infty} \prod_{u=1}^{t} \frac{\gamma_u}{\beta_u + \gamma_u} = 0, \tag{4.1}$$

if: i) $G(0) = 0$ with probability 1 and ii) $\Delta G(t) = U_t \prod_{u=1}^{t-1}(1 - U_u)$ for all $t \geq 1$, where $\{U_t : t \geq 1\}$ is a sequence of independent random variables such that $U_t \sim Beta(\beta_t, \gamma_t)$ for all integers $t \geq 1$. Condition (4.1) is both necessary and sufficient for a random function $G(t)$ satisfying points i) and ii) to be a cumulative distribution function with probability one. The beta-Stacy process prior is conjugate with respect to right-censored data, a property that makes it especially suitable in survival analysis applications. Moreover, if G is a discrete-time beta-Stacy process with parameters $\{(\beta_t, \gamma_t) : t \geq 1\}$, then the predictive distribution G^* of a new, yet unseen observation from G is determined by $\Delta G^*(t) = E[\Delta G(t)] = \frac{\beta_t}{\beta_t + \gamma_t} \prod_{u=1}^{t-1} \frac{\gamma_u}{\beta_u + \gamma_u}$, the probability that a new observation from G will be equal to t.

To generalize this approach to competing risks, we introduce the following definitions:

Definition 4.2.1. *A function $F : \{0, 1, 2, \ldots\} \times \{1, \ldots, k\} \to [0, 1]$, $k \geq 1$, is called a (discrete-time)* subdistribution function *if it is the joint distribution function of some random vector $(T, \delta) \in \{0, 1, 2, \ldots\} \times \{1, \ldots, k\}$: $F(t, c) = P(T \leq t, \delta = c)$ for all $t \geq 0$ and $c \in \{1, \ldots, k\}$. A* random subdistribution function *is defined as*

a stochastic process indexed by $\{0, 1, 2, \ldots\} \times \{1, \ldots, k\}$ *whose sample paths form a subdistribution function almost surely.*

Suppose now that T represents the time until one of k specific competing events occurs and that $\delta = 1, \ldots, k$ indicates the type of the occurring event. For instance, for $k = 2$, T may represent time from surgery for melanoma to death, while δ may represent the specific cause of death: $\delta = 1$ for melanoma-related mortality, $\delta = 2$ for death due to other causes. As before, if the distribution of (T, δ) is unknown, then in the Bayesian nonparametric framework it is assumed that, conditionally on some random subdistribution function F, (T, δ) is distributed according to F itself: $P(T \leq t, \delta = c|F) = F(t, c)$ for all $t \geq 0$ and $c = 1, \ldots, k$.

Remark 4.2.1. Conditionally on F, $\Delta F(t, c) = F(t, c) - F(t-1, c)$ is the probability of experiencing an event of type c at time t: $\Delta F(t, c) = P(T = t, \delta = c|F)$. Additionally, if $G(t) = \sum_{d=1}^{k} F(t, d)$, $\Delta G(t) = G(t) - G(t-1)$, and $V_{t,d} = \Delta F(t, d)/\Delta G(t)$, then: $G(t) = P(T \leq t|F)$ is the cumulative probability of experiencing an event by time t, $\Delta G(t) = P(T = t|F)$ is the probability of experiencing an event at time t, and $V_{t,c} = P(\delta = c|T = t, F)$ is the probability of experiencing an event of type c at time t given that some event occurs at time t. Moreover, it can be shown that $F(t, c) = \sum_{u=1}^{t} S(u-1)\Delta A_c(u)$, where $S(t) = 1-G(t)$ and $A_c(t) = \Delta F(t, c)/S(t-1)$ is the *cumulative hazard* of experiencing an event of type c by time t (Kalbfleisch and Prentice, 2002, Chapter 8).

To specify a suitable prior on the random subdistribution function F, we now introduce the subdistribution beta-Stacy process:

Definition 4.2.2. *Let* $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ *be a collection of* $(k+1)$*-dimensional vectors of positive real numbers satisfying the following condition:*

$$\lim_{t \to +\infty} \prod_{u=1}^{t} \frac{\alpha_{u,0}}{\sum_{d=0}^{k} \alpha_{u,d}} = 0. \tag{4.2}$$

A random subdistribution function F *is said to be a discrete-time* subdistribution beta-Stacy *process with parameters* $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ *if:*

1. $F(0, c) = 0$ *with probability 1 for all* $c = 1, \ldots, k$*;*

2. *for all* $c = 1, \ldots, k$ *and all* $t \geq 1$*,*

$$\Delta F(t, c) = W_{t,c} \prod_{u=1}^{t-1} \left(1 - \sum_{d=1}^{k} W_{u,d}\right), \tag{4.3}$$

with the convention that empty products are equal to 1*, and where* $\{W_t = (W_{t,0}, \ldots, W_{t,k}) : t \geq 1\}$ *is a sequence of independent random vectors such that for all* $t \geq 1$*,* $W_t \sim Dirichlet_{k+1}(\alpha_{t,0}, \ldots, \alpha_{t,k})$*.*

If in particular the $\alpha_{t,d}$ *are determined as*

$$\alpha_{t,c} = \omega_t \Delta F_0(t,c) \quad \textit{and} \quad \alpha_{t,0} = \omega_t \left(1 - \sum_{d=1}^{k} F_0(t,d)\right)$$

for some fixed subdistribution function F_0 *and sequence of positive real numbers* $(\omega_t : t \geq 1)$*, then we write* $F \sim s\mathcal{BS}(\omega, F_0)$*.*

Remark 4.2.2. In Section 4.3, Remark 4.3.1, it will be shown that condition (4.2) is both necessary and sufficient for a random function $F(t,c)$ satisfying points 1 and 2 of Definition 4.2.2 to be a subdistribution function with probability 1. This justifies the consideration of the subdistribution beta-Stacy process as a prior distribution on the space of subdistribution functions. Also note that if $F \sim s\mathcal{BS}(\omega, F_0)$, then condition (4.2) is automatically satisfied since $\sum_{d=0}^{k} \alpha_{t,d} = \omega_t(1 - \sum_{d=1}^{k} F_0(t-1,d))$ and so $\prod_{t=1}^{+\infty}[\alpha_{t,0}/\sum_{d=0}^{k} \alpha_{t,d}] = \lim_{t\to+\infty}(1 - \sum_{d=1}^{k} F_0(t,d)) = 0$, as $\lim_{t\to+\infty} \sum_{d=1}^{k} F_0(t,d) = 1$ (provided occurrence of at least one of the k events is inevitable).

The following lemma (which can be proven by taking expectations of Equation (4.3) and using the fact that the W_t are independent Dirichlet random vectors) characterizes the moments of the subdistribution beta-Stacy process.

Lemma 4.2.1. *Let* F *be a subdistribution beta-Stacy process with parameters* $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$*. Then*

$$E[\Delta F(t,c)] = \frac{\alpha_{t,c}}{\sum_{d=0}^{k} \alpha_{t,d}} \prod_{u=1}^{t-1} \frac{\alpha_{u,0}}{\sum_{d=0}^{k} \alpha_{u,d}}, \tag{4.4}$$

$$E[\Delta F(t,c)^2] = E[\Delta F(t,c)] \frac{1 + \alpha_{t,c}}{1 + \sum_{d=0}^{k} \alpha_{t,d}} \prod_{u=1}^{t-1} \frac{1 + \alpha_{u,0}}{1 + \sum_{d=0}^{k} \alpha_{u,d}} \tag{4.5}$$

for all $t \geq 1$ *and* $c = 1, \ldots, k$*.*

Remark 4.2.3. Using Theorem 2.5 of Ng et al. (2011) it is possible to show that the vector of random probabilities $(1 - \sum_{d=1}^{k} \Delta F(t,d),\ \Delta F(t,1),\ \ldots,\ \Delta F(t,k))$ is *completely neutral* in the sense of Connor and Mosimann (1969). Consequently, Equations (4.4) and (4.5) also follow from formulas (4) and (9) of Connor and Mosimann (1969).

Remark 4.2.4. Note that the previous Lemma 4.2.1 also characterizes the predictive distribution associated to a subdistribution beta-Stacy process: if F is a subdistribution beta-Stacy process with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$, then the predictive subdistribution function F^* of a new, yet unseen observation from F is determined by $\Delta F^*(t,d) = E[\Delta F(t,d)]$, which is given by Equation (4.4) ($\Delta F^*(t,d)$ is the probability that a new observation from F will be equal to (t,d)).

Remark 4.2.5. Let $F \sim s\mathcal{BS}(\omega, F_0)$. It can be shown from Equation (4.4) that $E[\Delta F(t,c)] = \Delta F_0(t,c)$ for all $t \geq 1$ and $c = 1, \ldots, k$, implying both that i) F is centered on F_0 and ii) F_0 is equal to the predictive distribution associated to F. From Equations (4.4) and (4.5) it can be further shown that $\mathrm{Var}\,(\Delta F(t,c))$ is a decreasing function of ω_t, with $\mathrm{Var}\,(\Delta F(t,c)) \to 0$ as $\omega_t \to +\infty$ and $\mathrm{Var}\,(\Delta F(t,c)) \to F_0(t,c)(1 - F_0(t,c))$ as $\omega_t \to 0$. Thus ω_t can be used to control the prior precision of the $s\mathcal{BS}(\omega, F_0)$ process.

4.3 Predictive characterisation

Muliere et al. (2000) described a predictive construction of the discrete-time beta-Stacy process by means of a reinforced urn process $\{Y_n : n \geq 0\}$ with state space $\{0, 1, 2, \ldots\}$. The urns of this process contain balls of only two colors, black and white (say), and reinforcement is performed by the addition of a single ball ($m = 1$). To intuitively describe this process, suppose that each patient in a series is observed from an initial time point until the onset of an event of interest. The process $\{Y_n : n \geq 0\}$ starts from $Y_0 = 0$, signifying the start of the observation for the first patient, and then evolves as follows: if $Y_n = t$ and a black ball is extracted, then the current patient does not experience the event at time t and $Y_{n+1} = t+1$; if instead a white ball is extracted, then the current patient experiences the event at time t and $Y_{n+1} = 0$, so the process is restarted to signify the start of the observation of a new patient. With this interpretation, the number T_n of states visited by $\{Y_n : n \geq 0\}$ between the $(n-1)$-th and n-th visits to the initial state 0 correspond to the time of event onset for the n-th patient. If the process $\{Y_n : n \geq 0\}$ is recurrent (so the times T_n are almost surely finite), a representation theorem for reinforced urn processes implies that the process $\{Y_n : n \geq 0\}$ is a mixture of Markov Chains. The corresponding mixing measure is such that the rows of the transition matrix are independent Dirichlet processes (Muliere et al., 2000, Theorem 2.16; see Ferguson,

1973 for the definition of a Dirichlet process). Using this representation, Muliere et al. (2000) showed that the sequence $\{T_n : n \geq 1\}$ is exchangeable and that there exists a random distribution function G such that i) conditionally on G, the times $T_1, T_2, \ldots$ are i.i.d. with common distribution function G, and ii) G is a beta-Stacy process (Muliere et al., 2000, Section 3).

In this section, we will generalize the predictive construction of Muliere et al. (2000) to yield a similar characterization of the subdistribution beta-Stacy process. To do so, consider a reinforced urn process $\{X_n : n \geq 0\}$ with state space $S = \{0, 1, 2, \ldots\} \times E$, set of colors $E = \{0, 1, \ldots, k\}$ $(k \geq 1)$, starting point $X_0 = (0, 0)$, and law of motion defined by $q((t, 0), c) = (t + 1, c)$ and $q((t, d), c) = (0, 0)$ for all for all integers $t \geq 0$ and $c, d = 0, 1, \ldots, k$, $d \neq 0$. Further suppose, for simplicity of presentation, that reinforcement is performed by the addition of a single ball $(m = 1)$ as before (but see Remark 4.3.2 below for the case where $m \in (0, +\infty)$). The initial composition of the urns is given as follows: i) $n_{(t,0)}(c) = \alpha_{t+1,c}$ for all integers $t \geq 0$ and $c = 0, 1, \ldots, k$; ii) $n_{(t,d)}(0) = 1$, $n_{(t,d)}(c) = 0$ for all integers $t \geq 0$ and $c, d = 1, \ldots, k$, $d \neq 0$. Now, define $\tau_0 = 0$ and $\tau_{n+1} = \inf\{t > \tau_n : X_t = (0, 0)\}$ for all integers $n \geq 0$. The process $\{X_n : n \geq 0\}$ is said to be *recurrent* if $P(\cap_{n=1}^{+\infty}\{\tau_n < +\infty\}) = 1$. Additionally, let $T((t, c)) = t$ and $D((t, c)) = c$ for all $(t, c) \in S$. For all $n \geq 1$, set $T_n = T(X_{\tau_n - 1})$, the length of the sequence of states between the $(n - 1)$-th and the n-th visits to the initial state $(0, 0)$, and $D_n = D(X_{\tau_n - 1})$, the color of the last ball extracted before the n-th visit to $(0, 0)$.

The process $\{X_n : n \geq 0\}$ can be interpreted as follows: a patient initially at risk of experiencing any of k possible outcomes is followed in time starting from time $t = 0$; at each time point t, the color of the extracted ball represents the status of the patient at the next time point $t + 1$; if a ball of color 0 is extracted, the patient remains at risk at the next time point; if instead a ball of color $c \in \{1, \ldots, k\}$ is extracted, then the patient will experience an outcome of type c at the next time point. The process returns to the initial state after such an occurrence to signify the arrival of a new patient. With this interpretation, the variable T_n represents the time at which the n-th patient experiences one of the k events under study, while D_n encodes the type of the realized outcome. These concepts are illustrated in Figure 4.1. Moreover, note that, although slightly different, the reinforced urn process used to construct the beta-Stacy process by Muliere et al. (2000) is essentially equivalent to the process $\{X_n : n \geq 0\}$ in the particular case where $k = 1$, with color 0 being

black and color 1 being white in the above description.

Continuing, in accordance with Diaconis and Freedman (1980) we say that the process $\{X_n : n \geq 0\}$ is *Markov exchangeable* if $P(X_0 = x_0, \ldots, X_n = x_n) = P(X_0 = y_0, \ldots, X_n = y_n)$ for all finite sequences $(x_0, \ldots, x_n)$ and $(y_0, \ldots, y_n)$ of elements of S such that i) $x_0 = y_0$ and ii) for any $s_1, s_2 \in S$, the number of transitions from s_1 to s_2 is the same in both sequences.

Lemma 4.3.1. *The process $\{X_n : n \geq 0\}$ is Markov exchangeable. Consequently, if $\{X_n : n \geq 0\}$ is recurrent, then it is also a mixture of Markov Chains with state space S. In other words, there exists a probability measure μ on the space $\mathcal{M}$ of all transition matrices on $S \times S$ and a $\mathcal{M}$-valued random element $\Pi \sim \mu$ such that for all $n \geq 1$ and all sequences $x_0, \ldots, x_n \in S$ with $x_0 = (0, 0)$,*

$$P(X_0 = x_0, \ldots, X_n = x_n | \Pi) = \prod_{i=0}^{n-1} \Pi(x_i, x_{i+1}),$$

where $\Pi(x, y)$ is the element on the x-row and y-th column of Π. Additionally, for each $x = (t, c) \in S$, let $\mathcal{N}_x(\cdot)$ be the measure on S (together with the Borel σ-algebra generated by the discrete topology) which gives mass $n_{(t,c)}(d)$ to $q((t, c), d)$ for all $d = 0, 1, \ldots, k$, and null mass to all other points in S. Then, the random probability measure $\Pi(x, \cdot)$ on S is a Dirichlet process with parameter measure $\mathcal{N}_x(\cdot)$.

Proof. The thesis follows immediately from Theorem 2.3 and 2.16 of Muliere et al. (2000) and Theorem 7 of Diaconis and Freedman (1980). □

Lemma 4.3.2. *The process $\{X_n : n \geq 0\}$ is recurrent if and only if $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ satisfies condition (4.2).*

Proof. First observe that

$$\begin{aligned}
P(\tau_1 = +\infty) &= \lim_{n \to +\infty} P(\tau_1 > n) \\
&= \lim_{n \to +\infty} P(X_0 = (0, 0), X_1 = (1, 0), \ldots, X_{n-1} = (n - 1, 0)) \\
&= \lim_{n \to +\infty} \prod_{t=0}^{n-1} \frac{n_{(t,0)}(0)}{\sum_{d=1}^{k} n_{(t,0)}(d)} \\
&= \lim_{n \to +\infty} \prod_{t=1}^{n} \frac{\alpha_{t,0}}{\sum_{d=1}^{k} \alpha_{t,d}}.
\end{aligned}$$

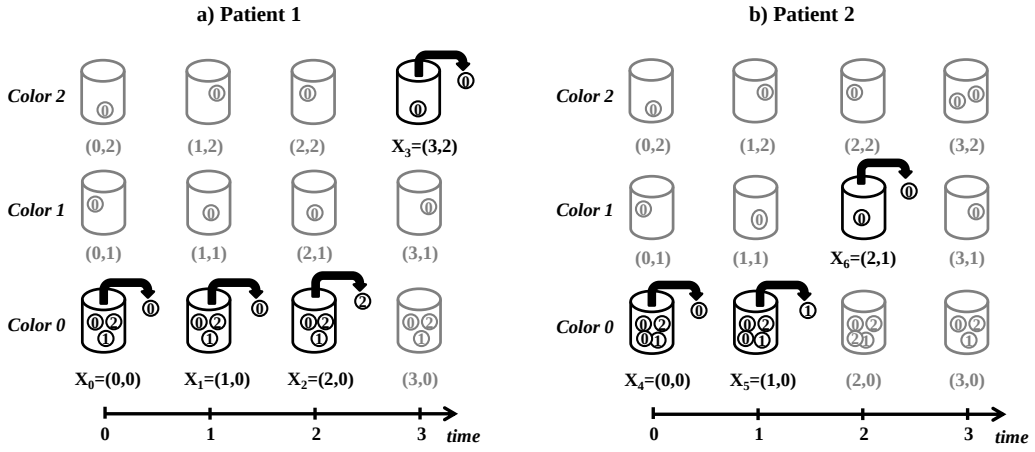


Figure 4.1: Illustration of the reinforced urn process characterizing the subdistribution beta-Stacy process assuming $k = 2$. In both panels, the horizontal axis measures the time since the last visit to the urn representing the state $(0, 0)$. The process starts from the $(0, 0)$ urn in Panel a, in which all urns are represented at their initial composition. In this example, balls of colors 0, 0, and 2 are successively extracted from the urns visited by the process, respectively at times 0, 1, and 2. At time 3 the process visits the $(3, 2)$ urn, from which only balls of color 0 can be extracted. The process then returns to the $(0, 0)$ urn and continues as shown in Panel b, where the composition of the urns has been updated by reinforcement. Suppose now that each visit to $(0, 0)$ represents the arrival of a new melanoma patient at the moment of surgery. If color 1 represents death due to melanoma and color 2 represents death due to other causes, then the sequence of urns visited in Panel a corresponds to the history of an individual (Patient 1) who dies of causes not related to melanoma after 3 time instants since surgery ($T_1 = 3$, $D_1 = 2$), while Panel b represents the history of a subsequently observed individual (Patient 2) who dies due to melanoma after 2 time instants since surgery ($T_2 = 2$, $D_2 = 1$).

Consequently, if $\{X_n : n \geq 0\}$ is recurrent, then $P(\tau_1 = \infty) = 0$ and so condition (4.2) must hold. Conversely, suppose that condition (4.2) is satisfied. Then $P(\tau_1 < +\infty) = 1$. By induction on $n \geq 1$, suppose that $P(\cap_{i=1}^{n}\{\tau_i < +\infty\}) = 1$. Then

$$P(\tau_{n+1} = +\infty) = \int_{\cap_{i=1}^{n}\{\tau_i<+\infty\}} P(\tau_{n+1} = +\infty|T_1, \ldots, T_n)dP.$$

Given $T_1, \ldots, T_n$, if $\tau_{n+1} = +\infty$ then the process must visit all states $(t, 0)$ with $t \geq 0$ starting from time τ_n. Since the states $(t, 0)$ for $t > L := \max(T_1, \ldots, T_n) + 1$ correspond to previously unvisited urns, the probability of this event is bounded above by

$$\lim_{n\to+\infty} \prod_{i=L}^{n} \frac{n_{(i,0)}(0)}{\sum_{d=1}^{k} n_{(i,0)}(d)} = \lim_{n\to+\infty} \prod_{i=L+1}^{n} \frac{\alpha_{i,0}}{\sum_{d=1}^{k} \alpha_{i,d}}.$$

Hence

$$P(\tau_{n+1} = +\infty) \leq \int_{\cap_{i=1}^{n}\{\tau_i<+\infty\}} \lim_{n\to+\infty} \prod_{i=L+1}^{n} \frac{\alpha_{i,0}}{\sum_{d=1}^{k} \alpha_{i,d}} dP = 0,$$

where the last equality follows from condition (4.2). Consequently,

$$P(\cap_{i=1}^{n+1}\{\tau_i < +\infty\}) = 1.$$

This argument shows that $P(\cap_{i=1}^{+\infty}\{\tau_i < +\infty\}) = 1$ and so the process must be recurrent, as needed. □

Theorem 4.3.1. *Suppose that the process $\{X_n : n \geq 0\}$ is recurrent. Then there exists a random subdistribution function F, such that, given F, the (T_n, D_n) are i.i.d. distributed according to F. Moreover, i) F is determined as a function of the random transition matrix Π from Lemma 4.3.1, and ii) F is a subdistribution beta-Stacy process with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$.*

Proof. Let Π be the random transition matrix on $S \times S$ provided by Lemma 4.3.1 and define $F(t, c) = P(T_1 \leq t, D_1 = c|\Pi)$, which is clearly a random subdistribution function. Moreover, for all $c = 1, \ldots, k$,

$$F(0, c) = P(T_1 = 0, D_1 = c|\Pi) \leq P(T(X_{\tau_1-1}) = 1|\Pi) = P(\tau_1 = 1|\Pi) = 0.$$

Instead, for all $c = 1, \ldots, k$ and all $t \geq 1$,

$$\begin{aligned}\Delta F(t,c) &= P(T_1 = t, D_1 = c|\Pi) \\ &= P(X_0 = (0,0), \ldots, X_{t-1} = (t-1,0), X_t = (t,c)|\Pi) \\ &= \Pi((t-1,0),(t,c)) \prod_{u=0}^{t-2} \Pi((u,0),(u+1,0)).\end{aligned}$$

Now, for all $t \geq 1$ and $d = 0, 1, \ldots, k$,

$$\mathcal{N}_{(t-1,0)}(\{(t,d)\}) = \mathcal{N}_{(t-1,0)}(\{q((t-1,0),d)\}) = n_{(t-1,0)}(d) = \alpha_{t,d}.$$

Then, from Lemma 4.3.1 again and from the properties of the Dirichlet process (Ferguson, 1973), for all $t \geq 1$, $\big(\Pi((t-1,0),(t,0)), \ldots, \Pi((t-1,0),(t,k))\big) \sim Dirichlet_{k+1}(\alpha_{t,0}, \ldots, \alpha_{t,k})$. Hence, Lemma 4.3.2 implies that F is subdistribution beta-Stacy with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$.

To show that, given F, the (T_n, D_n) are i.i.d. distributed according to F, it suffices to note that for all $(t_1, d_1), \ldots, (t_n, d_n) \in S$ such that $t_i \geq 1$ for all $i = 1, \ldots, n$, it holds that

$$\begin{aligned}&P\left((T_1, D_1) = (t_1, d_1), \ldots, (T_n, D_n) = (t_n, d_n)|\Pi\right) \\ &= \prod_{i=1}^{n} \left\{ \Pi((t_i - 1, 0), (t_i, d_i)) \prod_{t=0}^{t_i - 1} \Pi((t,0),(t+1,0)) \right\} \\ &= \prod_{i=1}^{n} \Delta F(t_i, d_i).\end{aligned}$$

Since F is a function of Π, this concludes the proof. □

Remark 4.3.1. Suppose that F is a random function satisfying points 1 and 2 of Definition 4.2.2. The proof of Theorem 4.3.1 also shows that, if condition (4.2) is satisfied, then F is a random subdistribution function. This is because condition (4.2) coincides with the recurrency condition in Lemma 4.3.2. Suppose instead that F is a subdistribution function with probability 1. Then $\tilde{F}(t,c) = E[F(t,c)]$ is a subdistribution function and

$$P(T_1 \leq t, D_1 = c) = \tilde{F}(t,c) = \frac{\alpha_{t,d}}{\sum_{c=0}^{k} \alpha_{t,c}} \prod_{u=1}^{t-1} \frac{\alpha_{u,0}}{\sum_{c=0}^{k} \alpha_{u,c}}$$

for all $t \geq 0$ and $c = 1, \ldots, k$. Hence it must be

$$\begin{aligned} 0 &= P(T_1 = +\infty) \\ &= \lim_{t \to +\infty} P(X_0 = (0,0), \ldots, X_t = (t,0)) \\ &= \lim_{t \to +\infty} \prod_{u=1}^{t} \frac{\alpha_{u,0}}{\sum_{c=0}^{k} \alpha_{u,c}}. \end{aligned}$$

Thus condition (4.2) must hold. Therefore, condition (4.2) is both necessary and sufficient for F to be a random subdistribution function, justifiying the claim anticipated in Remark 4.2.2.

Another immediate consequence of Theorem 4.3.1 is the following:

Corollary 4.3.1. *The sequence of random variables* $\{(T_n, D_n) : n \geq 1\}$ *induced by the reinforced urn process* $\{X_n : n \geq 0\}$ *is exchangeable.*

This fact could also have been proven directly through an argument similar to that at the end of Section 2 of Muliere et al. (2000). To elaborate, suppose that $\{Y_n : n \geq 0\}$ is a recurrent stochastic process with countable state space S and such that $X_0 = x_0 \in S$ with probability one. Then a x_0-*block* is defined as any finite sequence of states visited by process which begins from x_0 and ends at the state immediately preceding the successive visit to x_0. Diaconis and Freedman (1980) showed that if $\{Y_n : n \geq 0\}$ is also Markov exchangeable, then the sequence $\{B_n : n \geq 1\}$ of its x_0-blocks is exchangeable. Now, consider the reinforced urn process $\{Y_n : n \geq 0\}$ used by Muliere et al. (2000) for constructing the beta-Stacy process and described at the beginning of this section. This process is Markov exchangeable and so, under a recurrency condition, its sequence of 0-blocks $\{B_n : n \geq 1\}$ is exchangeable. Consequently, so must be the corresponding sequence of total survival times $\{T_n = f(B_n) : n \geq 1\}$, where $f(B)$ is the length of the 0-block B after excluding its initial element. Each 0-block B_n must have the form $(0, 1, \ldots, t)$ for some $t \geq 1$ and $f((0, 1, \ldots, t)) = t$ for all $t \geq 1$.

In our setting, it can easily be seen that the $(0,0)$-blocks of the reinforced urn process $\{X_n : n \geq 0\}$ introduced in this section are finite sequences of states of the form $((0,0), (1,0), \ldots, (t-1,0), (t,c))$ for some $t \geq 1$ and $c = 1, \ldots, k$. Any such $(0,0)$-block represents the entire observed history of an individual at risk of developing any one of the k considered competing risks. For example, the history of Patient

1 in Figure 4.1(a) is represented by the $(0,0)$-block $B_1 = ((0,0),(1,0),(2,0),(3,2))$, while that of Patient 2 in Figure 4.1(b) is represented by the $(0,0)$-block $B_2 = ((0,0),(1,0),(2,1))$. If $\{X_n : n \geq 0\}$ is recurrent, by Lemma 4.3.1 its sequence of $(0,0)$-blocks $\{B_n : n \geq 1\}$ is exchangeable. Hence, so must be the sequence $\{(T_n, D_n) = f(B_n) : n \geq 1\}$, as claimed, where $f(B)$ is the last state in the $(0,0)$-block B. For the example in Figure 4.1, $f(B_1) = (T_1, D_1) = (3,2)$ and $f(B_2) = (T_2, D_2) = (2,1)$.

Remark 4.3.2. Throughout this section, we have assumed for simplicity that each extracted ball is reinforced by only by single ball of the same color, i.e. $m = 1$. In general, a number $m > 0$ could be considered. It is possible to show (see for example Amerio et al., 2004 or Mezzetti et al., 2007) that Theorem 4.3.1 would still hold with F distributed according a subdistribution beta-Stacy process with parameters $\{(\alpha_{t,0}/m, \ldots, \alpha_{t,k}/m) : t \geq 1\}$. In particular, if $\alpha_{t,c} = \omega_t \Delta F_0(t,c)$ and $\alpha_{t,0} = \omega_t(1 - \sum_{d=1}^{k} F_0(t,d))$, then $F \sim s\mathcal{BS}(\omega/m, F_0)$. Hence, the number of balls m used for reinforcement can be used to control concentration of the prior around its mean.

4.4 Posterior distributions and censoring

Suppose that (T_i, D_i) is distributed according to some subdistribution function F and $T_i > 0$ with probability 1 for all $i = 1, \ldots, n$. If the value (T_i, D_i) can be potentially right-censored at the known time $c_i \in \{0, 1, 2, \ldots\} \cup \{+\infty\}$, then instead of observing the actual value (T_i, D_i) one is only able to observe (T_i^*, D_i^*), where $(T_i^*, D_i^*) = (T_i, D_i)$ if $T_i \leq c_i$ and $(T_i^*, D_i^*) = (c_i, 0)$ if $T_i > c_i$ (if $c_i = +\infty$, then (T_i, D_i) is not affected by censoring). The following theorem shows that the subdistribution beta-Stacy process has a useful conjugacy property even in presence of such right-censoring mechanism.

Theorem 4.4.1. *Suppose that* (T_1, D_1), ..., (T_n, D_n) *is an i.i.d. sample from a subdistribution function* F *distributed as a subdistribution beta-Stacy process with parameters* $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$. *If* (T_1, D_1), ..., (T_n, D_n) *are potentially right-censored at the known times* $c_1, \ldots, c_n$, *respectively, then the posterior distribution of* F *given* (T_1^*, D_1^*),..., (T_n^*, D_n^*) *is a subdistribution beta-Stacy with parameters* $\{(\alpha_{t,0}^*, \ldots, \alpha_{t,k}^*) : t \geq 1\}$, *where* $\alpha_{t,0}^* = \alpha_{t,0} + l_t + m_{t,0}$, $\alpha_{t,d}^* = \alpha_{t,d} + m_{t,d}$ *for all*

integers $t \geq 1$ *and for* $d = 1, \ldots, k$, *where*

$$l_t = \sum_{i=1}^{n} I\{T_i^* > t\}$$

and

$$m_{t,d} = \sum_{i=1}^{n} I\{T_i^* = t, D_i = d\}$$

for all $t \geq 1$ *and* $d = 0, 1, \ldots, k$.

Proof. To prove the thesis, it suffices it is true for $n = 1$, as the general case will then follow from an immediate induction argument. To do so, first note that, with reference to the renforced urn process $\{X_n : n \geq 0\}$ of Section 4.3, condition (4.2) implies that F can be seen as a function of some random transition matrix Π as in the proof of Theorem 4.3.1. Assume now that $(T_1^*, D_1^*) = (t, d)$ for some $t \geq 1$ and $d = 0, 1, \ldots, k$. Since observing (T_1^*, D_1^*) is equivalent to observing $X_0 = (0, 0), \ldots, X_{t-1} = (t-1, 0), X_t = (t, d)$, Corollary 2.21 of Muliere et al. (2000) implies that, conditionally on $(T_1^*, D_1^*) = (t, d)$, the rows of Π are independent and, for all $x \in S$, the parameter measure of the x-th row of Π assigns mass $n_{(0,0)}(0) + 1, \ldots, n_{(t-2,0)}(0)+1, n_{(t-1,0)}(d)+1$ to the states $(1, 0), \ldots, (t-1, 0), (t, d)$, respectively, and mass $n_{(t',d')}(c)$ to all other states $q((t', d'), c) \neq (1, 0), \ldots, (t-1, 0), (t, d)$ in S. Since $\alpha_{t,d} = n_{(t-1,0)}(d)$ for all $t \geq 1$ and $d = 0, 1, \ldots, k$, it can now be seen that, conditionally on (T_1^*, D_1^*), F must be subdistribution beta-Stacy with parameters $\{(\alpha_{t,0}^*, \ldots, \alpha_{t,k}^*) : t \geq 1\}$ defined by $\alpha_{t,0}^* = \alpha_{t,0} + I\{T_1^* > t\} + I\{T_1^* = t, D_1^* = 0\}$, $\alpha_{t,d}^* = \alpha_{t,d} + I\{T_1^* = t, D_1^* = d\}$ for all integers $t \geq 1$ for $d = 1, \ldots, k$. □

The following corollary is now a direct consequence of Equation (4.4) in Lemma 4.2.1.

Corollary 4.4.1. *The predictive distribution* $F^*(t, d)$ *of a new (non-censored) observation* (T_{n+1}, D_{n+1}) *from* F *having previously observed* $(T_1^*, D_1^*), \ldots, (T_n^*, D_n^*)$ *is determined by*

$$\begin{aligned}
\Delta F^*(t, d) &= P((T_{n+1}, D_{n+1}) = (t, d) | (T_1^*, D_1^*), \ldots, (T_n^*, D_n^*)) \\
&= E\left[\Delta F(t, d) | (T_1^*, D_1^*), \ldots, (T_n^*, D_n^*)\right] \\
&= \frac{\alpha_{t,d}^*}{\sum_{c=0}^{k} \alpha_{t,c}^*} \prod_{u=1}^{t-1} \frac{\alpha_{u,0}^*}{\sum_{c=0}^{k} \alpha_{u,c}^*}.
\end{aligned}$$

for all $t \geq 1$ *and* $d = 1, \ldots, k$.

The following result instead follows from Corollary 4.4.1 and Remark 4.2.5.

Corollary 4.4.2. *Assume that $F \sim s\mathcal{BS}(\omega, F_0)$ a priori. Then, the posterior distribution of F given the observed values of $(T_1^*, D_1^*), \ldots, (T_n^*, D_n^*)$ is $s\mathcal{BS}(\omega^*, F^*)$, where*

$$\begin{aligned}
F^*(t,c) &= \sum_{u=1}^{t} S^*(u-1)\Delta A_c^*(u), \\
A_c^*(t) &= \sum_{u=1}^{t} \frac{\omega_u \Delta F_0(u,c) + m_{u,c}}{\omega_u(1-\sum_{d=1}^{k} F_0(u-1,d)) + l_u + \sum_{d=0}^{k} m_{u,d}}, \\
S^*(t) &= \prod_{u=1}^{t} \left(1 - \frac{\omega_u \sum_{d=1}^{k} \Delta F_0(u,d) + \sum_{d=1}^{k} m_{u,d}}{\omega_u(1-\sum_{d=1}^{k} F_0(u-1,d)) + l_u + \sum_{d=0}^{k} m_{u,d}}\right),
\end{aligned}$$

and

$$\omega_t^* = \frac{\omega_t \left[1 - \sum_{d=1}^{k} F_0(t,d)\right] + l_t + m_{t,0}}{1 - \sum_{d=1}^{k} F^*(t,d)}.$$

Remark 4.4.1. As $\max_{u=1,\ldots,t}(\omega_u) \to 0$, $S^*(t)$ converges to the discrete-time Kaplan-Meier estimate

$$\widehat{S}(t) = \prod_{u=1}^{t}(1 - [\sum_{d=1}^{k} m_{u,d}]/[l_u + \sum_{d=0}^{k} m_{u,d}]),$$

while $A_c^*(t)$ converges to the Nelson-Aalen estimate

$$\widehat{A}_c(t) = \sum_{u=1}^{t} m_{u,c}/(l_u + \sum_{d=0}^{k} m_{u,d})$$

for all times $t \geq 1$ for which $\widehat{S}(t)$ and the $\widehat{A}_c(t)$ are defined (i.e. such that $l_t + \sum_{d=0}^{k} m_{t,d}] > 0$). All in all, $F^*(t,c)$, which coincides with the optimal Bayesian estimate of F under a squared-error loss, converges to

$$\widehat{F}(t,c) = \sum_{u=1}^{t} \widehat{S}(u-1)\Delta\widehat{A}_c(u),$$

the classical non-parametric estimate of $F(t,c)$ of Kalbfleisch and Prentice (2002, Chapter 8), for all times $t \geq 1$ for which this is defined. Conversely, if $\min_{u=1,\ldots,t}(\omega_u) \to +\infty$, then $S^*(t)$ converges to $1 - \sum_{d=1}^{k} F_0(t,d)$, $A_c(t)$ converges to the corresponding cumulative hazard of F_0, and therefore $F^*(t,c)$ converges to the prior mean $F_0(t,c)$ for all times $t \geq 1$ and $c = 1, \ldots, k$.

Remark 4.4.2. (Censored data likelihood) Given a sample (t_1^*, d_1^*), ..., (t_n^*, d_n^*) of censored observations from a subdistribution function $F(t, c)$, define $z_i = I\{d_i^* \neq 0\}$ for all $i = 1, \ldots, n$. It can then be shown that the likelihood function for F is

$$\begin{aligned} L(F) &= P\left((T_1^*, D_1^*) = (t_1^*, d_1^*), \ldots, (T_n^*, D_n^*) = (t_n^*, d_n^*)|F\right) \\ &= \prod_{i=1}^{n} \Delta F(t_i^*, d_i^*)^{z_i} \left[1 - \sum_{d=1}^{k} F(t_i^*, d)\right]^{1-z_i} . \end{aligned} \tag{4.6}$$

So far the censoring times $c_1, \ldots, c_n$ have been considered fixed and known. Theorem 4.4.1 however continues to hold also in the following more general setting in which censoring times are random: let the censored data be defined as $T_i^* = \min(T_i, C_i)$ and $D_i^* = I\{T_i \leq C_i\}$ for all $i = 1, \ldots, n$, where i) $C_1, \ldots, C_n$ are independent random variable with common distribution function $H(t)$, ii) conditional on F and H, $(T_1, D_1), \ldots, (T_n, D_n)$ and $C_1, \ldots, C_n$ are independent, and iii) F and H are a priori independent. Adapting the terminology of Heitjan and Rubin (1991; 1993), in this case the random censoring mechanism is said to be *ignorable*.

Theorem 4.4.2. *If censoring is random and ignorable and F is a priori a subdistribution beta-Stacy process, then the marginal likelihood for F is proportional to the likelihood $L(F)$ defined in Equation (4.6). Consequently, the posterior distribution of F given (T_1^*, D_1^*), ..., (T_n^*, D_n^*) is the same as that described in Theorem 4.4.1.*

Proof. The likelihood function for F and H given a sample (t_1^*, d_1^*),..., (t_n^*, d_n^*) of observations affected from ignorable random censoring is

$$\begin{aligned} L^*(F, H) &= P\left((T_1^*, D_1^*) = (t_1^*, d_1^*), \ldots, (T_n^*, D_n^*) = (t_n^*, d_n^*)|F, H\right) \\ &= L(F) \prod_{i=1}^{n} \Delta H(t_i^*)^{1-z_i} \left[1 - H(t_i^*)\right]^{z_i} \\ &= L(F) L^*(H), \end{aligned}$$

where L and the z_i are defined as in Equation 4.6. Therefore, the marginal likelihood for F is $L^{\text{marginal}}(F) = L(F) E_H[L^*(H)] \propto L(F)$, where the constant of proportionality only depends on the data and $E_H[\cdot]$ represents expectation with respect to the prior distribution of H. As a consequence, the posterior distribution of F can be computed ignoring the randomness in the censoring times $C_1, \ldots, C_n$ by considering their observed values as fixed and their unobserved values as fixed to $+\infty$. Hence,

if F is a priori a subdistribution beta-Stacy process, then its posterior distribution is the same as in Theorem 4.4.1. □

Remark 4.4.3. The update-rule of Theorem 4.4.1 could be shown to hold under even more general censoring mechanisms. In fact, the marginal likelihood for F remains proportional to $L(F)$ as long as i) the distribution H of censoring times is independent of F and ii) censoring only depends on the past and outside variation (Kalbfleisch and Prentice, 2002).

4.5 Relation with other prior processes

4.5.1 Relation with the beta-Stacy process

By construction, the subdistribution beta-Stacy process can be regarded as a direct generalization of the beta-Stacy process. In fact, the two processes are linked with each other, as highlighted by the following theorem:

Theorem 4.5.1. *A random subdistribution function F is a discrete-time subdistribution beta-Stacy process with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ if and only if i) $G(t) = \sum_{d=1}^{k} F(t,d)$ is a discrete-time beta-Stacy process with parameters $\{(\sum_{d=1}^{k} \alpha_{t,d}, \alpha_{t,0}) : t \geq 1\}$ and ii) $\Delta F(t,c) = V_{t,c} \Delta G(t)$ for all $t \geq 1$ and $c = 1, \ldots, k$, where $\{V_t = (V_{t,1}, \ldots, V_{t,k}) : t \geq 1\}$ is a sequence of independent random vectors independent of G and such that $V_t \sim Dirichlet_k(\alpha_{t,1}, \ldots, \alpha_{t,k})$ for all $t \geq 1$ (where, if $k = 1$, we let the distribution $Dirichlet_1(\alpha_{t,1})$ be the point mass at 1).*

Proof. Before proceeding, first observe that $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ satisfies the recurrency condition of Equation (4.2) if and only if

$$\{(\sum_{d=1}^{k} \alpha_{t,d}, \alpha_{t,0}) : t \geq 1\}$$

satisfies the recurrency condition for the beta-Stacy process, i.e. Equation (4.1) with $\beta_t = \sum_{d=1}^{k} \alpha_{t,d}$ and $\gamma_t = \alpha_{t,0}$.

Now, to prove the "if" part of the thesis, suppose that the random subdistribution function F is subdistribution beta-Stacy with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$. Let $G(t) = \sum_{d=1}^{k} F(t,d)$ for all integers $t \geq 0$ and define $U_t = \sum_{d=1}^{k} W_{t,k}$

and $V_{t,c} = W_{t,c}/U_t$ for all $t \geq 1$ and $c = 1, \ldots, k$. From these definitions it is easy to check that $\Delta F(t,c) = V_{t,c}\Delta G(t)$ for all $t \geq 1$. Additionally, by standard properties of the Dirichlet distribution (Sivazlian, 1981, Propriety 1B) and by the independence of the W_t, $\{U_t : t \geq 1\}$ is a sequence of independent random variables such that $U_t \sim Beta(\sum_{d=1}^{k} \alpha_{t,k}, \alpha_{t,0})$. Moreover, from Theorem 2.5 of Ng et al. (2011) it follows that $\{V_t = (V_{t,1}, \ldots, V_{t,k}) : t \geq 1\}$ is a sequence of independent random vectors such that $V_t \sim Dirichlet_k(\alpha_{t,1}, \ldots, \alpha_{t,k})$ for all $t \geq 1$, independently of $\{U_t : t \geq 1\}$. Moreover, $G(0) = 0$ with probability one because $G(0) = \sum_{d=1}^{k} F(0,d)$ and $F(0,d) = 0$ with probability 1 for all $d = 1, \ldots, k$. Continuing, since $\Delta G(t) = \sum_{d=1}^{k} \Delta F(t,d)$ for all $t \geq 1$, it follows that

$$\Delta G(t) = \sum_{d=1}^{k} \left\{ W_{t,d} \prod_{u=1}^{t-1} \left(1 - \sum_{c=1}^{k} W_{u,c} \right) \right\} = U_t \prod_{u=1}^{t-1} (1 - U_u)$$

for all $t \geq 1$. Thus G is a beta-Stacy process with parameters

$$\{(\sum_{d=1}^{k} \alpha_{t,k}, \alpha_{t,0}) : t \geq 1\}.$$

To prove the "only if" part of the thesis, suppose instead that G is a beta-Stacy process with parameters $\{(\sum_{d=1}^{k} \alpha_{t,d}, \alpha_{t,0}) : t \geq 1\}$ and that $\{V_t = (V_{t,1}, \ldots, V_{t,k}) : t \geq 1\}$ is a sequence of independent random vectors satisfying conditions (a) and (b). Since $0 = G(0) = \sum_{d=1}^{k} F(0,d)$ with probability 1, and since it must also be $F(t,d) \geq 0$ with probability 1 for all t and d, it follows that $F(0,d) = 0$ with probability 1 for all d. To continue, define $W_t = (W_{t,0}, W_{t,1}, \ldots, W_{t,k}) = (1 - U_t, U_t V_{t,1}, \ldots, U_t V_{t,k})$ for all $t \geq 1$. It can be seen that the W_t are independent. Moreover, since U_t and V_t are independent and since $1 - U_t \sim Beta(\alpha_{t,0}, \sum_{d=1}^{k} \alpha_{t,d})$, from Theorem 2.2 of Ng et al. (2011), it follows that W_t has the same distribution as

$$\left(Y_{t,0}, Y_{t,1}(1 - Y_{t,0}), \ldots, Y_{t,k-1} \prod_{d=0}^{k-2} (1 - Y_{t,d}), \prod_{d=0}^{k-1} (1 - Y_{t,d}) \right)$$

where the $Y_{t,c}$ are independent random variables with

$$Y_{t,c} \sim Beta\left(\alpha_{t,c}, \sum_{d=c+1}^{k} \alpha_{t,d} \right)$$

for all $c = 0, \ldots, k$. Again from Theorem 2.2 of Ng et al. (2011) it thus follows that $W_t \sim Dirichlet_{k+1}(\alpha_{t,0}, \ldots, \alpha_{t,k})$ for all $t \geq 1$. Since $\Delta F(t,c) = V_{t,c}\Delta G(t)$ for all $t \geq 1$, from the definition of a beta-Stacy process it now follows that

$$\Delta F(t,d) = V_{t,d}U_t \prod_{u=1}^{t-1}\left(1 - \sum_{c=1}^{k} V_{t,c}U_u\right) = W_{t,d}\prod_{u=1}^{t-1}\left(1 - \sum_{c=1}^{k} W_{t,c}\right).$$

Hence, F is subdistribution beta-Stacy with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$. □

4.5.2 Relation with the beta process

Suppose $A(t) = (A_1(t), \ldots, A_k(t))$ collects the cumulative hazards of the subdistribution function $F(t,c)$ and let $\Delta A(t) = (\Delta A_1(t), \ldots, \Delta A_k(t))$, $A_0(t) = \sum_{d=1}^{k} A_d(t)$. Then, following Hjort (Hjort, 1990, Section 2), a discrete time beta-process prior for non-homogeneous Markov Chains with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$ could be specified for $A(t)$ by independently letting $(1 - \Delta A_0(t), \Delta A_1(t), \ldots, \Delta A_k(t))$ have a $Dirichlet(\alpha_{t,0}, \ldots, \alpha_{t,k})$ distribution for all $t \geq 1$. In such case, from Definition 4.2.2 it would follow that F is subdistribution beta-Stacy with the same set of parameters. The converse is also true, since if F is subdistribution beta-Stacy then it can be easily seen from Definition 4.2.2 that $(1 - \Delta A_0(t), \Delta A_1(t), \ldots, \Delta A_k(t)) = (W_{t,0}, W_{t,1}, \ldots, W_{t,k})$. Thus, if interest is in the subdistribution function $F(t,c)$ itself, one should consider the subdistribution beta-Stacy process, whereas if interest is in the cumulative hazards $A(t)$, one should consider the beta process for non-homogeneous Markov Chains. This equivalence parallels an analogous relation between the usual beta-Stacy and beta processes (Walker and Muliere, 1997).

4.5.3 Relation with the beta-Dirichlet process

The subdistribution beta-Stacy process is also related to the discrete-time version of the *beta-Dirichlet* process, a generalization of Hjort's beta process prior (Hjort, 1990) introduced by Kim and Gray (2012). The cumulative hazards $\{A(t) : t \geq 1\}$ are said to be a beta-Dirichlet process with parameters $\{(\beta_{t,1}, \beta_{t,2}, \gamma_{t,1}, \ldots, \gamma_{t,k}) : t \geq 1\}$ if i) the $\Delta A(t)$ are independent, ii) $\Delta A_0(t) \sim Beta(\beta_{t,1}, \beta_{t,2})$ for all $t \geq 1$, and iii) $\Delta A(t)/\Delta A_0(t) \sim Dirichlet_k(\gamma_{t,1}, \ldots, \gamma_{t,k})$ independently of $\Delta A_0(t)$ for all $t \geq 1$. From Definition 4.2.2 it is clear that if $F(t,c)$ is subdistribution beta-Stacy with parameters $\{(\alpha_{t,0}, \ldots, \alpha_{t,k}) : t \geq 1\}$, then from $(1 - \Delta A_0(t), \Delta A_1(t), \ldots, \Delta A_k(t)) =$

$(W_{t,0}, W_{t,1}, \ldots, W_{t,k})$ and Theorem 2.5 of Ng et al. (2011), then the corresponding cumulative hazards $A(t)$ must be beta-Dirichlet with parameters $\beta_{t,1} = \sum_{d=1}^{k} \alpha_{t,d}$, $\beta_{t,2} = \alpha_{t,0}$, and $\gamma_{t,d} = \alpha_{t,d}$ for all $d = 1, \ldots, k$ and $t \geq 1$. The converse is not true unless $\beta_{t,1} = \sum_{d=1}^{k} \gamma_{t,d}$ for all $t \geq 1$.

4.6 Nonparametric cumulative incidence regression

In this section, we will illustrate a subdistribution beta-Stacy regression approach for competing risks. We consider data represented by a sample of possibly-right censored discrete survival times and cause-of-failure indicators (t_1^*, d_1^*), ..., (t_n^*, d_n^*). Each observation (t_i^*, d_i^*) is associated with a known vector w_i of predictors. We assume that, as described in the Introduction, the time axis has been discretized according to some fixed partition $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ representing the measurement scale of event times. Hence, $(t_i^*, d_i^*) = (t, d)$ for some $d = 1, \ldots, k$ if an event of type d has been observed in the time interval $(\tau_{t-1}, \tau_t]$. Instead, $(t_i^*, d_i^*) = (t, 0)$ if no event has been observed during $(\tau_{t-1}, \tau_t]$ and censoring took place in the same interval.

Our starting point is the assumption that individual observations are exchangeable within each level of the predictor variables w_i. This leads us to consider a hierarchical modelling approach akin to that adopted by Lindley and Smith (1972), Antoniak (1974), and Cifarelli and Regazzini (1978). In this approach, the observations $(t_1^*, d_1^*), \ldots, (t_n^*, d_n^*)$ are assumed to be independent, each generated by a corresponding subdistribution function $F(t, c; w_i)$ under some censoring mechanism (as described in Section 4.4). Then a joint prior distribution is assigned to all the $F(t, c; w_i)$ for $i = 1, \ldots, n$. In the usual parametric approach these would simply be assigned a specific functional form $F_0(t, c|\theta; w_i)$ by letting $F(t, c; w_i) = F_0(t, c|\theta; w_i)$ for all $i = 1, \ldots, n$, then assigning a prior distribution to the common parameter vector θ. Consequently, in the parametric approach the only source of uncertainty is that on the value of θ and not on the functional form of $F_0(t, c|\theta; w_i)$. Thus, to incorporate this uncertainty in the model and gain more flexibility in representing the shape of the $F(t, c; w_i)$, we instead adopt a nonparametric perspective (Müller and Mitra, 2013; Hjort et al., 2010; Phadia, 2013; Ghosal and van der Vaart, 2017). Specifically, we consider the following modelling approach: first, a specific functional

form $F_0(t, c|\theta; w_i)$ is chosen; second, letting $w_{(1)}, \ldots, w_{(L)}$ denote the distinct values of $w_1, \ldots, w_n$, the subdistribution functions $F(\cdot; w_{(i)})$ are assumed to be independent and distributed as $F(\cdot; w_{(i)}) \sim s\mathcal{BS}(\omega(\theta, w_{(i)}), F_0(\cdot|\theta, w_{(i)}))$ for all $i = 1, \ldots, L$, where $\omega(\theta, w_{(i)}) = (\omega_t(\theta, w_{(i)}))_{t\geq 1}$; finally, the parameter vector θ is assigned its own prior distribution. The use of weights $\omega_t(\theta, w_{(i)}))$ dependent on the time location t, the covariate values $w_{(i)}$, and the parameter θ allows the local control of the uncertainty on the functional form of $F_0(\cdot|\theta, w_{(i)})$ by determining the concentration of the subdistribution beta-Stacy prior around the increments $\Delta F_0(t, c|\theta, w_{(i)})$, $c = 1, \ldots, k$ (c.f. Remark 4.2.5). Additionally, the presence of the parameter vector θ in the model allows borrowing of information across different covariate levels $w_{(i)}$, as in Cifarelli and Regazzini (1978), Muliere and Petrone (1993), and Mira and Petrone (1996).

Many options are available in the literature for specifying the functional form of the centering parametric subdistribution $F_0(t, c|\theta, w_i)$ when adopting our modelling approach. In general, following Larson and Dinse (1985), a useful strategy consists in starting from the decomposition

$$F_0(t, c|\theta, w_i) = F_0^{(1)}(c|\theta_1, w_i)F_0^{(2)}(t|\theta_2, c, w_i),$$

and then separately modelling the probability $F_0^{(1)}(c|\theta_1, w_i)$ of observing a failure of type c and the conditional distribution $F_0^{(2)}(t|\theta_2, c, w_i)$ of event times given the specific failure type c. For example, $F_0^{(1)}(c|\theta, w_i)$ can be specified to be any model for multinomial responses, such as the familiar multinomial logistic regression model

$$\begin{aligned} F_0^{(1)}(c|\theta_1, w_i) &= \frac{\exp(w_i' b_c)}{1 + \sum_{d=1}^{k-1} \exp(w_i' b_d)}, \\ F_0^{(1)}(k|\theta_1, w_i) &= \frac{1}{1 + \sum_{d=1}^{k-1} \exp(w_i' b_d)}, \end{aligned} \tag{4.7}$$

where $c = 1, \ldots, k-1$ and $\theta_1 = (b_1, \ldots, b_{c-1})$ (Agresti, 2003, Chapter 7). The time-to-event distribution $F_0^{(2)}(t|\theta_2, c, w_i)$ can instead be specified by discretizing a continuous-time distribution (e.g. Weibull or log-normal) with cumulative distribution function $G_0(\cdot|\theta_2, c, w)$ by letting $F_0^{(2)}(t|\theta_2, c, w_i) = G_0(\tau_t|\theta_2, c, w_i)$. For example, in the Weibull case one could let

$$G_0(t|\theta_2, c, w_i) = 1 - \exp(-t^{u_c} \exp(w_i' v_c)) \tag{4.8}$$

and $\theta_2 = (v_1, \ldots, v_k, u_1, \ldots, u_k)$, yielding a discrete-time version of the common parametric Weibull regression model (Aalen et al., 2008, Chapter 5). Alternatively, $F_0^{(2)}(t|\theta_2, c, w_i)$ may be specified as the Grouped Cox model (Kalbfleisch and Prentice, 2002, Section 2.4.2), the logistic regression model of Cox (1972), or other discrete-time models (Schmid et al., 2016; Tutz and Schmid, 2016; Berger and Schmid, 2017).

The weights $\omega_t(\theta, w_{(i)})$ can be calibrated in order to account for the degree of uncertainty attached to the chosen parametric functional form of $F_0(\cdot|\theta, w_i)$. In fact, by Remark 4.2.5, conditionally on θ as the weights increase the prior $s\mathcal{BS}(\omega, F_0(\cdot|\theta, w_{(i)}))$ becomes more concentrated on
$F_0(\cdot|\theta, w_{(i)}))$, giving more importance to the parametric component of the model. To exploit this fact, we consider weights ω_t defined as $\omega_t(\theta, w_{(i)}) = \omega_{0,t}(\theta, w_{(i)})/m$, where

$$\omega_{0,t}(\theta, w_{(i)}) = \frac{\tau_t - \tau_{t-1}}{\sum_{d=1}^{k} F_0(\tau_t, d|\theta, w_{(i)}) - \sum_{d=1}^{k} F_0(\tau_{t-1}, d|\theta, w_{(i)})}.$$

Extending the approach of Rigat and Muliere (2012), this choice allows the model to rely more on its parametric component over the times where observations are less likely to be available (high $\omega_{0,t}$), whereas it allows for more flexibility over the times where most data is expected (low $\omega_{0,t}$). The parameter m can be further used to control the standard deviations

$$\sigma_m(t, c; w_{(i)}) = \sqrt{\mathrm{Var}\left(\Delta F(t, c|w_{(i)}) - \Delta F_0(t, c|\theta, w_{(i)})\right)}$$

(computed from the joint distribution of $F(t, c|w_{(i)})$ and θ), which together measure how much the subdistribution function $F(\cdot; w_{(i)})$ can deviate from the parametric model $F_0(\cdot|\theta, w_{(i)})$. More precisely, by Remark 4.2.5, for all fixed $t \geq 1$ and $c = 1, \ldots, k$, is a decreasing function of m. In particular, $\sigma_m(t, c; w_{(i)}) \to 0$ for all t and c as $m \to 0$, implying that for $m \approx 0$ the model becomes essentially equivalent to the centering parametric model. Conversely, $\sigma_m(t, c; w_{(i)})$ increases to its maximum value

$$\sigma_\infty(t, c; w_{(i)}) = \sqrt{E[\Delta F_0(t, c|\theta, w_{(i)})(1 - \Delta F_0(t, c|\theta, w_{(i)}))]} \tag{4.9}$$

for all t and c as $m \to +\infty$. Hence, for large m the model becomes more flexible and is allowed to deviate more freely from the centering parametric model.

Remark 4.6.1. Conditionally on θ, the predictive structure of such model can be characterized as by associating an urn system like that described in Section 4.3 to

each distinct value of $w_{(i)}$. The initial composition of these urns is determined by

$$\alpha_{t,c}(\theta, w_{(i)}) = \omega_{0,t}(\theta, w_{(i)})\Delta F_0(t, c|\theta, w_{(i)})$$

and

$$\alpha_{t,0}(\theta, w_{(i)}) = \omega_{0,t}(\theta, w_{(i)}) \left(1 - \sum_{d=1}^{k} F_0(t, d|\theta, w_{(i)})\right).$$

If each extracted ball is reinforced by m similar balls, then by Theorem 4.3.1 and Remark 4.3.2, the distributions associated to the same value of $w_{(i)}$ are independent and each distributed according to some

$$F(\cdot|w_{(i)}) \sim s\mathcal{BS}((\omega_t(\theta, w_{(i)}))_{t\geq 1}, F_0(\cdot|\theta, w_{(i)})),$$

where $\omega_t(\theta, w_{(i)}) = \omega_{0,t}(\theta, w_{(i)})/m$ as above.

4.6.1 Sampling from the posterior distribution

To fix notations, let $t^* = (t_1^*, \ldots, t_n^*)$, $d^* = (d_1^*, \ldots, d_n^*)$, $w = (w_1, \ldots, w_n)$, and $\mathcal{F} = (F(\cdot; w_{(1)}), \ldots, F(\cdot; w_{(L)}))$. Also, let $n_j = \sum_{i=1}^{n} I\left\{w_i = w_{(j)}\right\}$ and for all $i = 1, \ldots, n_j$ let $(t_{j,1}^*, d_{j,1}^*)$, $\ldots$, $(t_{j,n_j}^*, d_{j,n_j}^*)$ be the set of observations corresponding to the value $w_{(j)}$. Lastly, let $z_{j,i} = I\left\{d_{j,i} \neq 0\right\}$ for all possible j and i. Finally, let $t_j^* = (t_{j,i}^* : i = 1, \ldots, n_j)$ and $d_j^* = (d_{j,i}^* : i = 1, \ldots, n_j)$ for all $j = 1, \ldots, L$.

Theorem 4.6.1. *Assuming ignorable right censoring, the marginal likelihood of θ is*

$$P(t^*, d^*|\theta, w) = \prod_{j=1}^{L}\prod_{i=1}^{n_j} \Delta F_{j,i-1}^*(t_{j,i}^*, d_{j,i}^*|\theta, w_{(j)})^{z_{j,i}} S_{j,i-1}^*(t_{j,i}^*|\theta, w_{(j)})^{1-z_{j,i}},$$

where:

$$S_{j,i-1}^*(t_{j,i}^*|\theta, w_{(j)}) = 1 - \sum_{d=1}^{k} F_{j,i-1}^*(t_{j,i}^*, d|\theta, w_{(j)}),$$

$F_{j,i}^*(t, d|\theta, w_{(j)})$ is the predictive distribution of a new observation from $F(\cdot|w_{(j)})$ given $(t_{j,1}^*, d_{j,1}^*)$, $\ldots$, $(t_{j,i}^*, d_{j,i}^*)$, obtained from Corollaries 4.4.1 and 4.4.2, and

$$F_{j,0}^*(t, d|\theta, w_{(j)}) = F_0(t, d|\theta, w_{(j)}).$$

Proof. First assume that censoring is fixed. In this case, the marginal likelihood of θ can be obtained from conditional likelihood

$$\begin{aligned} P(t^*, d^*|\theta, w, \mathcal{F}) &= \prod_{j=1}^{L} P(t_j^*, d_j^*|\theta, w_{(j)}, F(\cdot; w_{(j)})) \\ &= \prod_{j=1}^{L} P(t_j^*, d_j^*|F(\cdot; w_{(j)})) \end{aligned}$$

by taking its expectation with respect to the distribution of $\mathcal{F}$ conditional on θ. Since the $F(\cdot; w_{(j)})$ are independent conditionally on θ, the marginal likelihood is thus

$$\begin{aligned} P(t^*, d^*|\theta, w) &= \prod_{j=1}^{L} P(t_j^*, d_j^*|\theta, w_{(j)}) \\ &= \prod_{j=1}^{L} \prod_{i=1}^{n_j} P(T_{j,i}^* = t_{j,i}^*, D_{j,i}^* = d_{j,i}^*|t_{j,h}^*, d_{j,h}^*, h < i; \theta, w_{(j)}), \end{aligned}$$

where $P(T_{j,i}^* = t_{j,i}^*, D_{j,i}^* = d_{j,i}^*|t_{j,h}^*, d_{j,h}^*, h < i; \theta, w_{(j)})$ is the conditional predictive distribution of $(t_{j,i}^*, d_{j,i}^*)$ given all $(t_{j,h}^*, d_{j,h}^*)$ with $h < i$ and θ. If $z_{j,i}^* = 1$, this can be derived from Corollaries 4.4.2 and 4.4.1 and it is equal to $\Delta F_{j,i-1}^*(t_{j,i}^*, d_{j,i}^*|\theta)$. If instead $z_{j,i}^* = 0$, then this is equal to

$$P(T_{j,i} > t_{j,i}^*|(T_{j,h}^*, D_{j,h}^*) = (t_{j,h}^*, d_{j,h}^*), h < i; \theta, w_{(j)}) = S_{j,i-1}^*(t_{j,i}^*|\theta, w_{(j)}).$$

This justifies the thesis if censoring is fixed. By similar arguments as those in Section 4.4, the same likelihood can be assumed to hold also in presence of ignorable censoring, as needed. □

Using the above result, the joint posterior distribution $P(\mathcal{F}, \theta|t^*, d^*, w)$ of $\mathcal{F}$ and θ can be obtained as

$$P(\mathcal{F}, \theta|t^*, d^*, w) \propto P(\theta)P(t^*, d^*|\theta, w) \prod_{j=1}^{L} P_j(F(\cdot; w_{(j)})|\theta, w), \tag{4.10}$$

where $P(\theta)$ represents the prior distribution of θ (which is independent of w) and the term $P_j(F(\cdot; w_{(j)})|\theta, w)$ represents the posterior distribution of $F(\cdot; w_{(j)}) \sim s\mathcal{BS}(\omega, F_0(\cdot|\theta, w_{(j)}))$ obtained (for fixed θ) from the data $\mathcal{D}_j = \{(t_i^*, d_i^*) : w_i =$

$w_{(j)}, i = 1, \ldots, n\}$ using the update rule described in Theorem 4.4.1. Now, although the posterior distribution for θ is not available for exact sampling, Equation (4.10) suggests the use of a Markov Chain Monte Carlo strategy such as the following to perform approximate posterior inferences. First, a sample $\{\theta_i\}_{i=1}^S$ from the marginal posterior distribution of θ is obtained, after discarding an appropriate number of burn-in iterations, via a Random Walk Metropolis-Hastings algorithm (Robert and Casella, 2004, Section 7.5). A multivariate Gaussian distribution can be considered after the reparametrization induced by a logarithmic transformation of each shape parameter u_c (to account for their positive support). Second, having obtained a sample $\{\theta_i\}_{i=1}^S$ as just described, the conditional posterior distribution of $F(\cdot; w_{(j)})$ given θ_i and the data $\mathcal{D}_j$ is obtained by direct simulation for all $i = 1, \ldots, S$ and $j = 1, \ldots, L$. Specifically, the parameters of the conditional posterior distribution $P_j(F(\cdot; w_{(j)})|\theta, w)$ of $F(\cdot; w_{(j)})$ given θ_i and $\mathcal{D}_j$ are obtained using Theorem 4.4.1. Then a sample $F_i(\cdot; w_{(j)})$ from $P_j(F(\cdot; w_{(j)})|\theta, w)$ is obtained using Definition 4.2.2 by sampling from the relevant Dirichlet distributions. The sample $\{(\theta_i, F_i(\cdot; w_{(1)}), \ldots, F_i(\cdot; w_{(L)}))\}_{i=1}^S$ so obtained then represents a sample from the joint posterior distribution of Equation (4.10).

4.6.2 Estimating the predictive distributions

Let T_{n+1} and D_{n+1} be the unknown uncensored survival time and type of realized outcome, respectively, for a new individual with covariate profile w_{n+1}. The objective is to estimate the predictive distribution of (T_{n+1}, D_{n+1}) given the data (t_1^*, d_1^*), ..., (t_n^*, d_n^*). We distinguish two cases: i) $w_{n+1} = w_{(j)}$ for some $j = 1, \ldots, L$, and ii) $w_{n+1} \neq w_{(1)}, \ldots, w_{(L)}$. In the first case, simply obtain a sample $\{F_i(\cdot; w_{n+1}) = F_i(\cdot; w_{(j)})\}_{i=1}^S$ from the posterior distribution of $F(\cdot; w_{(j)})$ using the output of the procedure described above. The predictive distribution of (T_{n+1}, D_{n+1}) is then estimated as $S^{-1} \sum_{i=1}^S F_i(\cdot; w_{n+1})$. In the second case it is still possible to estimate the predictive distribution of (T_{n+1}, D_{n+1}) by recycling the sample $\{\theta_i\}_{i=1}^S$. Specifically, for each θ_i, $F_i(\cdot; w_{n+1})$ is simulated directly from the $s\mathcal{BS}(\omega, F_0(\cdot|\theta_i, w_{n+1}))$ distribution. The predictive distribution of (T_{n+1}, D_{n+1}) is then estimated as the average of the sampled subdistribution functions, as before.

4.7 Application: analysis of the melanoma dataset

4.7.1 Data description and analysis objectives

To illustrate our modelling approach, we analyse data collected by
Drzewiecki et al. (1980) on 205 stage I melanoma patients who underwent surgical excision of the tumor during 1962-1977 at the Odense University Hospital, Denmark. This dataset (which includes only data for those 205 patients for which an histological examination was carried out, out of the 225 originally participating in the study) has been previously used to illustrate several survival analysis methods (Andersen et al., 2012, Example I.3.1) and is freely available online as part of the *timereg* R library (Scheike and Zhang, 2011). Each considered patient was followed from the date of surgery to the time of death for melanoma (event of type 1), death due to other causes (event of type 2), or censoring (e.g. study drop-out or end of the study, defined at the end of 1977). Event times are only known discretized at the day level, so that $\tau_t = t$ for all $t \geq 0$ can be assumed. (the time-interval $(\tau_{t-1}, \tau_t]$ represents the t-th day of follow-up). In summary, 126 (61%) of the study participants were women and 79 (39%) were men. Overall, a total of 57 (28%) patients died due to melanoma during follow-up, while 14 (7%) died due to other causes, overall accumulating 441,324 person-days of follow-up (maximum follow-up: men, 4,492 days; women, 5,565 days). Using these data, we implement a competing-risks regression model to assess the long-term prognosis of melanoma patients following surgical excision of the tumor with respect to the risk of death due to melanoma. In doing so, we account for death due to other causes as a competing event and consider gender as a potential predictor.

4.7.2 Model specification and prior distributions

We consider a regression model specified as explained in Section 4.6. For illustration, we specify the centering parametric model $F_0(t, c|\theta, w_i)$ by consider the multinomial logistic model (4.7) for $F_0^{(1)}(c|\theta_1, w_i)$ and the discrete Weibull regression model (4.8) for $F_0^{(2)}(t|\theta_2, c, w_i)$, as these correspond to models widely used in applications. In these models, for all subjects $w_i = (w_{i,1}, w_{i,2})$ includes an intercept term ($w_{i,1} = 1$) and the indicator variable for gender ($w_{i,2} = 0$ for women, $w_{i,2} = 1$ for men). Consequently, in the notations of Section 4.6, $\theta_1 = (b_1)$, where $b_1 = (b_{1,1}, b_{1,2})$ is the vector

of the two regression coefficients in model (4.7) ($b_{1,1}$ for the intercept, $b_{1,2}$ for the gender indicator). Instead, $\theta_2 = (v_1, v_2, u_1, u_2)$, where $v_c = (v_{c,1}, v_{c,2})$ is the vector of the two regression coefficients for the cause-specific Weibull regression model (4.8) for $c = 1, 2$ ($v_{c,1}$ for the intercept, $v_{c,2}$ for the gender indicator), while $u_1, u_2 > 0$ are the two corresponding shape parameters. We assign independent prior distributions to all parameter as follows. Noting that Drzewiecki et al. (1980) estimated that the overall 10-years survival probability was about 50% (estimated via the Kaplan-Meier method) in a previous analysis of a larger dataset, we calibrate the priors for v_1 and v_2 in such a way so as to center the curves $F_0^{(2)}(t|\theta_2, c, w_i)$ around a model with a median survival of 3,650 days. To do so, we assigned $N(\log(-\log(0.50)/3,650), 1)$ priors to $v_{1,1}$ and $v_{2,1}$, and $N(0, 1)$ to $v_{1,2}$ and $v_{2,2}$. We assign a $N(0, 1)$ prior distributions to $b_{1,1}$, $b_{1,2}$ and a gamma distribution $Gamma(g_1, g_2)$ with shape parameter $g_1 = 11$ and rate parameter $g_2 = 10$ distribution to u_1 and u_2 (thus centering the corresponding Weibull distributions on an exponential model). Numerical simulations reported in the Appendix Subsection 4.9.1 suggest that these choices yield a fairly diffuse prior distribution for the subdistribution function of the model. As a sensitivity analysis, in the Appendix Subsection 4.9.3, we report the results obtained from a similar model but considering a discrete log-normal distribution for the centering parametric subdistribution function.

4.7.3 Calibrating the prior concentrations

To illustrate the behaviour of our model as m varies, Figure 4.2 shows, for the prior distributions specified in the previous section, the values of the prior standard deviations $\sigma_m(t, c; w_{(i)})$ for increasing values of m, computed by simulating from the priors described in the previous setting and focusing on the subdistribution of death due to melanoma ($c = 1$) among women ($w_{(i)} = (1, 0)$). Qualitatively identical results (data not shown) can be obtained for death due to other causes ($c = 2$) or men ($w_{(i)} = (1, 1)$). From these results show how, for small m (e.g. $m = 1$ in Figure 4.2), the nonparametric prior $s\mathcal{BS}(\omega(\theta, w_{(i)}), F_0(\cdot|\theta, w_{(i)}))$ for $F(t, c; w_{(i)})$ is practically fully concentrated on its parametric component (as $\sigma_m(t, c; w_{(i)}) \approx 0$ for most $t \geq 1$). This implies that for small m the subdistribution $F(t, c; w_{(i)})$ will tend to be almost equal to the parametric centering subdistribution $F_0(t, c|\theta; w_{(i)})$ a priori. However, as m increases, so does $\sigma_m(t, c; w_{(i)})$, representing increasing

levels of prior uncertainty on the functional form of $F(t, c; w_{(i)})$. For sufficiently large values of m (e.g. $m = 10^5$ in Figure 4.2), the $\sigma_m(t, c; w_{(i)})$ achieve values close to their upper bound $\sigma_\infty(t, c; w_{(i)})$, which represents a situation of maximum uncertainty on the functional form of $F(t, c; w_{(i)})$. Regardless of the value of m, the $\sigma_m(t, c; w_{(i)})$ decrease as $t \geq 1$ increases, showing how the parametric model $F_0(t, c|\theta; w_{(i)})$ is given more and more weight in determining the form of $F(t, c; w_{(i)})$ over the later portions of follow-up (i.e. over times where less data is expected a priori). These observations both illustrate the consideration of Section 4.6 but also suggest that plots like Figure 4.2 may be useful in practice to calibrate the prior concentration.

4.7.4 Posterior analysis

Posterior inference was performed by a Random Walk Metropolis-Hastings algorithm with a multivariate Gaussian proposal distribution as suggested in Section 4.6, by means of the *MCMCpack* R package (Martin et al., 2011). The proposal distribution was centered at the current sampled value, with a proposal covariance matrix equal to the negative inverse Hessian matrix of the log-posterior distribution, evaluated at the posterior mode and scaled by $(2.4)^2/d$, where d is the dimension of θ, as suggested by Gelman et al. (2013, Section 12.2). To improve mixing, all predictors were standardized before running the algorithm. The parameter vector was initialized with the corresponding value obtained by numerically maximizing the log-posterior distribution. In all cases, the Metropolis-Hastings algorithm was run for a total of 26000 iterations: the first 1000 were discarded as burn-in, while the remaining 25000 were thinned by retaining only one generated sample every 25 iterations. The trace plots of generated Markov Chain Monte Carlo chains did not raise any issue of non-convergence according to both Geweke's test (Geweke, 1992) and visual inspection (data not shown). Additionally, the obtained posterior distributions were found to be much more concentrated than the considered prior distributions, as shown in the on-line Appendix Subsection 4.9.1.

Figure 4.3 shows the posterior predictive distributions, i.e. the posterior expectations of the subdistribution functions $F(t, 1; w_{(i)})$, for death due to melanoma among men (panel a) and women (panel b), for $m = 10^0$, 10^3, or 10^6. For comparison, Figure 4.3 also reports i) the estimates obtained from the classical Kalbfleish-Prentice

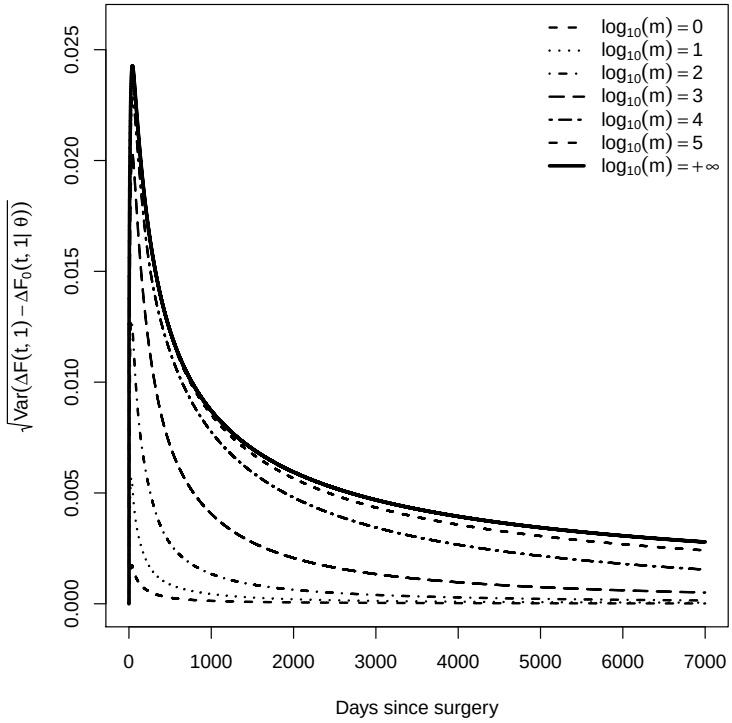
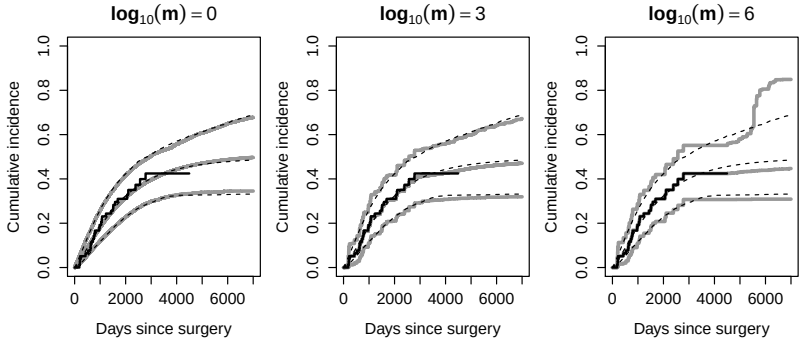
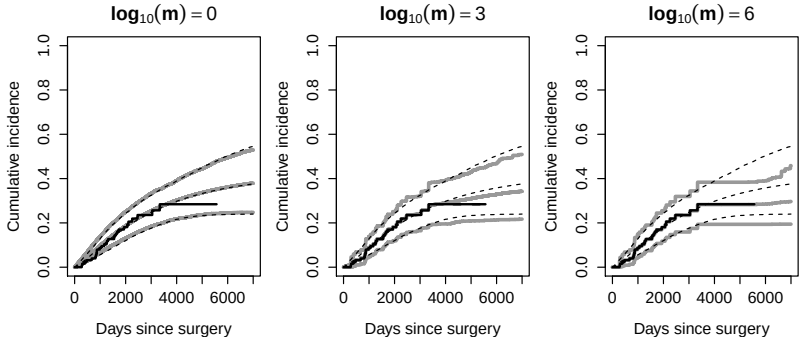


Figure 4.2: Prior standard deviations $\sigma_m(t, c; w_{(i)}) = \sqrt{\text{Var}\left(\Delta F(t, c|w_{(i)}) - \Delta F_0(t, c|\theta, w_{(i)})\right)}$ corresponding to the model of Section 4.7.2 for $m = 10^0, 10^1, \ldots, 10^5$, together with its upper bound $\sigma_\infty(t, c; w_{(i)})$ from Equation 4.9. Results are for the subdistribution of death due to melanoma ($c = 1$) among women ($w_{(i)} = (1, 0)$). The quantity $\sigma_m(t, c; w_{(i)})$ measures the concentration of the subdistribution beta-Stacy prior for the subdistribution function $F(t, c; w_{(i)})$ around the parametric subdistribution $F_0(t, c|\theta, w_{(i)})$. Values of $\sigma_m(t, c; w_{(i)}) \approx 0$ signify that $F(t, c; w_{(i)}) \approx F_0(t, c|\theta, w_{(i)})$ with high priori probability, while increasing values of $\sigma_m(t, c; w_{(i)}) > 0$ signify that functional forms different than $F_0(t, c|\theta, w_{(i)})$ are more likely a priori.



(a) Cumulative incidence of death due to melanoma, men.



(b) Cumulative incidence of death due to melanoma, women.

Figure 4.3: Posterior summaries for the cumulative incidence of death due to melanoma among (a) men and (b) women, i.e. posterior summaries for the subdistribution functions $F(t, 1; w_{(i)})$ for the model of Section 4.7 with (a) $w_{(i)} = (1, 1)$ and (b) $w_{(i)} = (1, 0)$, computed for reinforcement parameters $m = 10^0$, 10^3, and 10^6. Solid black lines: Kalbfleish-Prentice classical estimators. Solid gray lines: posterior means of the subdistribution function, i.e. posterior predictive distributions, with upper and lower 95% pointwise credibility limits. Dashed black lines, posterior means and 95% pointwise credibility limits for the multinomial-Weibull model of Section 4.6.

estimator and ii) the posterior estimates obtained from the centering multinomial-Weibull parametric model $F_0(t, c|\theta, w_i)$ of Section 4.7.2 (using the same parametric prior distributions for comparability). In general, the results are compatible with the observation that men are subject to a higher risk of death due to melanoma than women (Thörn et al., 1994). Additionally, from Figure 4.3 it is apparent how the classical estimators have a limited usefulness for evaluating long-term prognosis, as these are undefined beyond the range of the observed data. On the other hand, by relying more on its parametric component, our subdistribution beta-Stacy model can provide an extrapolated risk estimate. Risk extrapolations could also be obtained from the centering parametric model, but these would require absolute confidence in its assumed functional form. Oppositely, our model may deviate more or less flexibly from the centering parametric model according to the chosen value of m. In fact, the results obtained from the subdistribution beta-Stacy model for $m = 1$ are essentially equivalent to those obtained from centering parametric model. However, as m increases the subdistribution beta-Stacy predictive distributions better approximate the classical estimators of the subdistribution function. For $m = 10^6$ it can also be seen that the posterior variance of the subdistribution function may be very large beyond the range of the observed data, as seen in Figure 4.3 for men (panel a), i.e. the group that required the most extrapolation for computing the predictive distribution over the considered time period. This behaviour is consistent with the observations of Section 4.7.3: for large m the model allows more uncertainty on the functional form of the centering model.

Additional results are provided in the Appendix Subsections 4.9.2 and 4.9.3. Specifically, in Appendix 4.9.2 we report the results of graphical posterior predictive checks for the goodness of fit of our model, in the style of Gelman et al. (2013, Section 6.3). These checks do not raise any concern regarding the fit of our model. In Appendix 4.9.3, we report the results obtained in the sensitivity analysis based on the discrete log-normal model. The corresponding results are essentially equivalent to the ones obtained here.

4.7.5 Simulation study

To further explore how much our model can adapt to deviations from the corresponding centering parametric functional form, we conducted a simulation study as

follows. First, on the basis of the melanoma data of Section 4.7.1, we computed the maximum likelihood estimates $\widehat{\theta} = (\widehat{b}_{1,1}, \widehat{v}_{1,1}, \widehat{v}_{2,1}, \widehat{u}_1, \widehat{u}_2)$ for the multinomial-Weibull model $F_0(t, c|\theta)$ of Section 4.7.2, ignoring covariates for simplicity by including only an intercept term as predictor ($w_{(i)} \equiv 1$). We thus obtained $\widehat{b}_{1,1} = -0.640$, $\widehat{v}_{1,1} = -11.927$, $\widehat{v}_{2,1} = -7.244$, $\widehat{u}_1 = 1.597$, and $\widehat{u}_2 = 0.639$. Second, we generated 100 datasets of sample size $n = 100$, $n = 500$, and $n = 1000$ by simulating event times and event types from the subdistribution function $F_0(t, c|\widehat{\theta})$, with a fixed censoring time at 7000 days since surgery. Third, in each simulated dataset, we implemented two Bayesian models: i) a multinomial-Weibull parametric model akin to that in Section 4.7.2 but where all Weibull shape parameters where fixed as $u_1, u_2 \equiv 1$ (all other prior distributions taken as in Section 4.7.2); this corresponds to incorrectly modelling the event times as exponentially distributed given the type of occurring event, incompatibly with the data-generating mechanism; ii) a nonparametric subdistribution beta-Stacy model centered on the parametric model of point (i) for all values of $m = 10^0, 10^3, 10^6$. Fourth and last, for each replicated dataset we computed the Kolmogorov-Smirnov distance $\max_{t\in[0,7000]} \left|\widehat{F}(t, c) - F_0(t, c|\widehat{\theta})\right|$ between the fixed data-generating subdistribution function $F_0(t, c|\widehat{\theta})$ and its posterior estimate $\widehat{F}(t, c)$, obtained from model (i) or (ii).

Figure 4.4 reports the distribution of Kolmogorov-Smirnov distances obtained in the described simulation study. As expected, for all sample size the misspecified parametric model produces the highest median Kolmogorov-Smirnov distances between the posterior estimates of the subdistribution function and the true data-generating subdistribution function. In agreement with previous observations, for $m = 1$ the nonparametric subdistribution beta-Stacy model tends to agree with its parametric component, producing similar results as those obtained from the parametric model. For increasing m, however, the subdistribution beta-Stacy model attains a greater flexibility to deviate from its misspecified parametric centering model and adapt to the data, thus producing lower Kolmogorov-Smirnov distances. This phenomenon, which is consistent with the observations of Section 4.7.3, is evident for all considered sample sizes, but especially for $n = 1000$. For this sample size, even the subdistribution beta-Stacy model with $m = 1$ is associated with a lower median Kolmogorov-Smirnov distance from the data-generating model than its centering parametric model: despite a large prior weight was assigned to the centering mispecified model, the sample size was large enough for the subdistribution

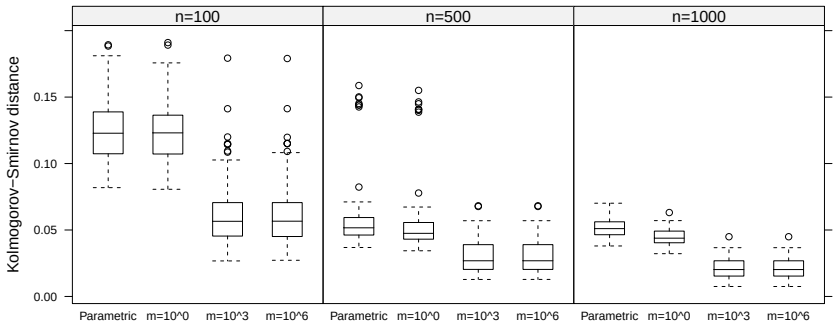


Figure 4.4: Box-plots reporting the distributions of Kolmogorov-Smirnov distances between the data-generating subdistribution function and its posterior estimates generated in the simulation study of Section 4.7.5 for sample sizes $n = 100$, $n = 500$, or $n = 1000$ (considering 100 simulated datasets per sample size). Kolmogorov-Smirnov distances are shown separately for the parametric model described in Section 4.7.5 and the nonparametric subdistribution beta-Stacy model, centered on the same parametric model, for reinforcement parameters of $m = 10^0, 10^3, 10^6$.

beta-Stacy model to be more driven by its nonparametric component and adapt to the data.

4.8 Concluding remarks

In this paper we introduced a novel stochastic process, the subdistribution beta-Stacy process, useful for the Bayesian nonparametric regression analysis of competing risks data. We showed how the subdistribution beta-Stacy process is completely characterized from a specific predictive structure, which we described in terms of the urn-based reinforced stochastic process of Muliere et al. (2000). The practical value of similar reinforced stochastic processes is that they potentially allow to undertake Bayesian predictive inference without explicit knowledge of the prior. That is, as noted by Muliere et al. (2003), they allow to update the predictive distributions from past information from a sequence of exchangeable observable without necessarily being able to compute the underlying de Finetti measure of the observations, i.e. the prior. In this paper, we were actually able to characterize the prior, which we identified as the subdistribution beta-Stacy. Although this process may have been

defined solely in terms of a sequence of independent Dirichlet random vectors (as in our Definition 2.2), the corresponding predictive construction still greatly simplifies the understanding of its properties.

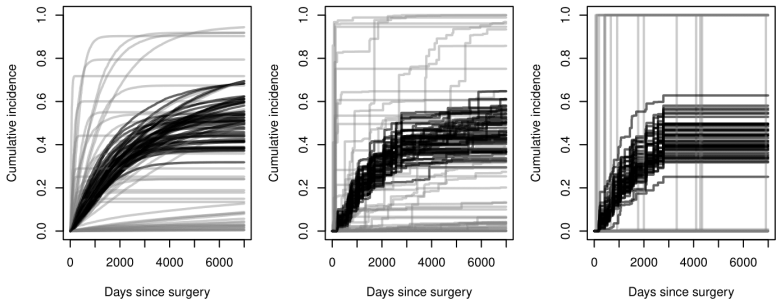
In this paper we also proposed a Bayesian nonparametric approach for competing risks regression based on the subdistribution beta-Stacy process. This provides several advantages with respect to other available techniques when making predictions in presence of competing risks. For instance, classical nonparametric estimators are typically undefined beyond the last observation time if this is censored, thus limiting their usefulness when making predictions. Parametric models circumvent this issue by assuming a specific functional form for the subdistribution function, thus providing risk extrapolations at the cost of more rigidity when adapting to the data. Conversely, by balancing both a nonparametric and a parametric component, our approach allows risk extrapolations beyond the range of the observed data without losing flexibility in adapting to available data. Additionally, contrary to most approaches available in the literature, our does not require the proportional hazards assumption, thus increasing its flexibility in capturing complex patterns of subdistribution functions when making predictions.

To conclude, we remark that more general reinforced urn processes may lead to interesting novel approaches for performing Bayesian inference in presence of competing risks or more general settings. In fact, we are currently investigating the following possible generalizations. First, akin as in Muliere et al. (2006), each extracted ball may be reinforced by a positive random number of new similar balls. This may depend on both the color of the ball extracted from the urn and the state represented by the urn itself. Such construction could be useful to represent allow uncertainty on the strength of belief to be granted to the initial composition of the urns. Second, a continuous-time version of the subdistribution beta-Stacy process could be obtained by embedding a discrete-time reinforced urn process like that of Section 4.3 into a reinforced continuous-time arrival process (representing the predictive distribution of the event times) as in the approach of Muliere et al. (2003). Third, the reinforced urn process considered in Section 4.3 could be generalized to characterize a process prior on the space of transition kernels of a Markovian multistate process, with application to the analysis of event-history data (Aalen et al., 2008). The conceptual difficulty here is that such processes may not be recurrent, complicating the use of representation theorems like those of Diaconis

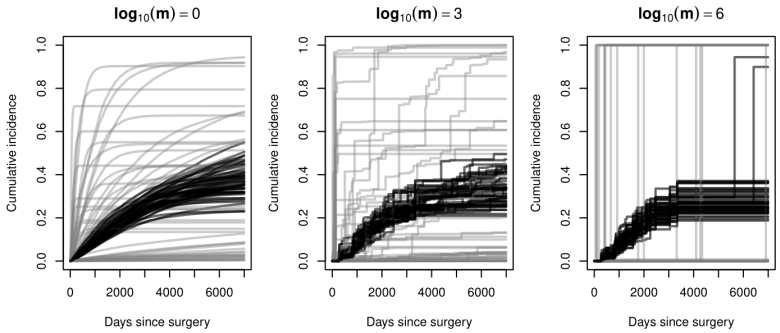
and Freedman (1980).

4.9 Appendix

4.9.1 Prior and posterior samples



(a) Cumulative incidence of death due to melanoma, men.



(b) Cumulative incidence of death due to melanoma, women.

Figure 4.5: Plots of 50 samples from the prior and posterior distribution for the cumulative incidence of death due to melanoma among (a) men and (b) women, i.e. prior and posterior subdistribution functions $F(t, 1; w_{(i)})$ for the model of Section 7 of the main paper with (a) $w_{(i)} = (1, 1)$ and (b) $w_{(i)} = (1, 0)$. Solid black lines: posterior samples. Solid gray lines: prior samples.

Figure 4.5 shows the graphs of 50 samples from the prior and posterior distribution for the cumulative incidence of death due to melanoma among (a) men and (b) women. The simulated curves show how the prior distribution for the subdistribution function is much more diffuse than the posterior distribution.

4.9.2 Posterior predictive checks

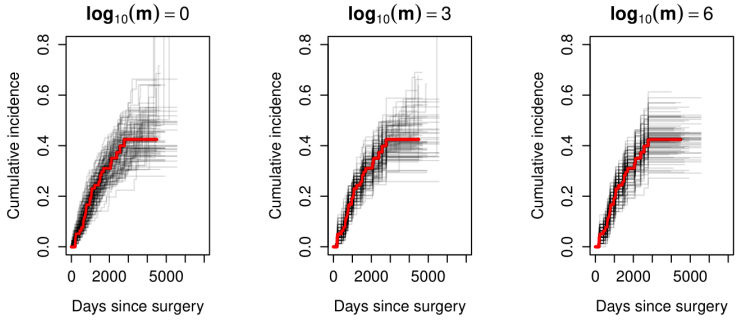
To assess the fit of the model of Section 7, we implemented some graphical posterior predictive checks. The idea is to display replicated data simulated from the posterior predictive distributions of the model alongside the original data and to look for systematic discrepancies between real and simulated datasets. If the model fits, then replicated data generated under the model should look similar to observed data. (Gelman et al., 2013, Section 6.3). To implement this, first we simulate replicated uncensored data from the posterior predictive distributions obtained from the model of Section 7 for $m = 10^0$, 10^3, and 10^6. Second, we apply censoring by simulating random censoring times from the empirical distribution of the observed censored observations. Third, we plot the classical Kalbfleish-Prentice estimator of the subdistribution function on the replicated censored data and compare it with that obtained from the original data. Results are reported in Figure 4.6. No alarming issue of lack of fit is raised from these results for the model of Section 7 for the considered values of m.

4.9.3 Analysis based on the discrete log-normal model

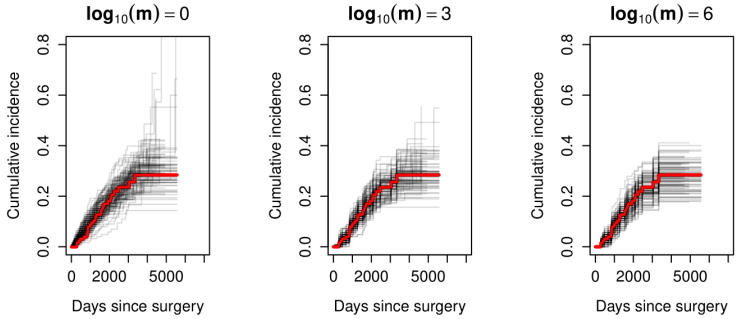
As a sensitivity analysis, we consider a modification of the model of Section 7 by considering a discrete log-normal distribution for the centering parametric subdistribution function $F_0(t, c|\theta, w_i)$. In more detail, we specify the centering parametric model $F_0(t, c|\theta, w_i)$ by consider a multinomial logistic model for $F_0^{(1)}(c|\theta_1, w_i)$ and a discrete log-normal regression model, obtained by letting $F_0^{(2)}(t|\theta_2, c, w_i) = G_0(\tau_t|\theta_2, c, w_i)$, where

$$G_0(t|\theta_2, c, w_i) = LogN(t; w_i' v_c, u_c), \tag{4.11}$$

where $LogN(t; \mu, \sigma)$ is the cumulative distribution function of a log-normal distribution with location parameter μ and scale parameter σ. Similarly as in the model of Section 7, for all subjects $w_i = (w_{i,1}, w_{i,2})$ includes an intercept term ($w_{i,1} = 1$) and the indicator variable for gender ($w_{i,2} = 0$ for women, $w_{i,2} = 1$ for men). Consequently, in the notations of Section 7, $\theta_1 = (b_1)$, where $b_1 = (b_{1,1}, b_{1,2})$ is the vector of the two regression coefficients in the multinomial logistic model model ($b_{1,1}$ for the intercept, $b_{1,2}$ for the gender indicator). Instead, $\theta_2 = (v_1, v_2, u_1, u_2)$, where $v_c = (v_{c,1}, v_{c,2})$ is the vector of the two regression coefficients for $c = 1, 2$ ($v_{c,1}$ for the intercept, $v_{c,2}$ for the gender indicator), while $u_1, u_2 > 0$ are the two corresponding



(a) Cumulative incidence of death due to melanoma, men.



(b) Cumulative incidence of death due to melanoma, women.

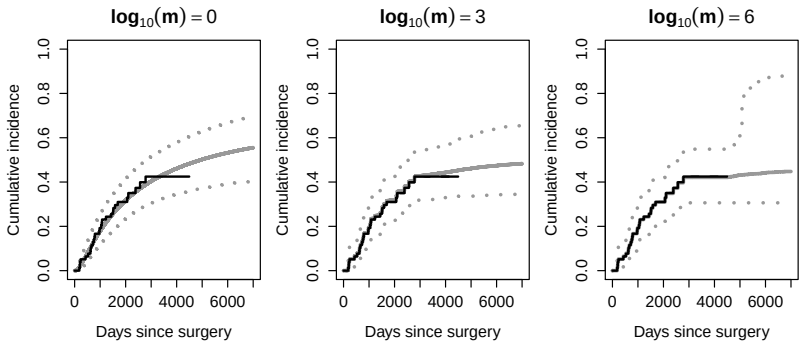
Figure 4.6: Plots of 100 replicates of the Kalbfleish-Prentice estimator of the sub-distribution function for death due to melanoma obtained by simulating replicated datasets from the posterior predictive distributions of the model of Section 7 and on the original dataset. Panel a: men. Panel b: women. Black solid curves: replicated datasets. Red solid curve: original dataset.

scale parameters. As also done in Section 7, we assign independent prior distributions to all parameter as follows. Noting that Drzewiecki et al. (1980) estimated that the overall 10-years survival probability was about 50% (estimated via the Kaplan-Meier method) in a previous analysis of a larger dataset, we calibrate the priors for v_1 and v_2 in such a way so as to center the curves $F_0^{(2)}(t|\theta_2, c, w_i)$ around a model with a median survival of 3,650 days. To do so, we assigned $N(\log(3,650), 1)$ priors to $v_{1,1}$ and $v_{2,1}$, and $N(0,1)$ to $v_{1,2}$ and $v_{2,2}$. We assign a $N(0,1)$ prior distributions to $b_{1,1}$, $b_{1,2}$ and a gamma distribution $Gamma(g_1, g_2)$ with shape parameter $g_1 = 2$ and rate parameter $g_2 = 1$ distribution to u_1 and u_2. Numerical simulations suggest that these choices yield a fairly diffuse prior distribution for the subdistribution function of the model (data not shown).

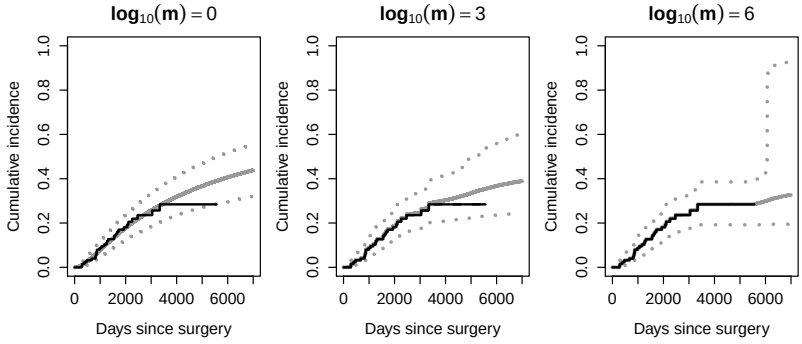
Figure 4.7 shows the obtained posterior predictive distributions, i.e. the posterior expectations of the subdistribution functions $F(t, 1; w_{(i)})$, for death due to melanoma among men (panel a) and women (panel b), letting $m = 10^0$, 10^3, or 10^6. For comparison, Figure 4.7 also reports the classical nonparametric estimators of the subdistribution functions. The results are qualitatively similar to those obtained in Section 7.

4.10 Available code

I developed R functions to implement the Weibull-SBS and lognormal-SBS regression models described in Sections 4.6 and Appendix Section 4.9.3, respectively. Computations are performed according to the strategy described in Sub-section 4.6.1. The random walk Metropolis-Hastings step of the algorithm is implemented using the `MCMCmetrop1R` function in the R library `MCMCpack` Martin et al. (2011). The functions are available at `https://github.com/andreaarfe/subdistribution-beta-stacy`, with R code to reproduce the analyses of Sections 4.7 and 4.9.



(a) Cumulative incidence of death due to melanoma, men.



(b) Cumulative incidence of death due to melanoma, women.

Figure 4.7: Posterior summaries for the cumulative incidence of death due to melanoma, obtained using a discrete log-normal centering subdistribution function, among (a) men and (b) women, i.e. posterior summaries for the subdistribution functions $F(t, 1; w_{(i)})$ for the model of Section 7 with (a) $w_{(i)} = (1, 1)$ and (b) $w_{(i)} = (1, 0)$. Solid dark lines: Kalbfleish-Prentice classical estimators, computed for reinforcement parameters $m = 10^0$, 10^3, and 10^6. Solid gray lines: posterior means of the subdistribution function, i.e. posterior predictive distributions. Dotted gray lines: upper and lower 95% pointwise credibility limits.

Chapter 5

The semi-Markov beta-Stacy process: a Bayesian non-parametric prior for semi-Markov processes

With Stefano Peluso and Pietro Muliere.
Manuscript under review.
ArXiv manuscript: `https://arxiv.org/abs/1812.00260`

5.1 Introduction

Discrete-time Semi-Markov processes generalize Markov chains by allowing the *holding times*, the times spent in each visited state, to have arbitrary distributions other than the geometric (Çinlar, 1969). In this paper, we address how to perform inferences and predictions for these processes from a Bayesian non-parametric perspective.

Because of their flexibility, discrete-time semi-Markov processes are used to predict many phenomena that evolve through a sequence of discrete states. Applications include time-series and longitudinal data analysis (Bulla and Bulla, 2006), survival analysis and reliability (Barbu et al., 2004; Mitchell et al., 2011), finance and actuarial sciences (Janssen and Manca, 2007), and biology (Barbu and Limnios,

2009).

Despite their usefulness in applications, and in contrast with their continuous-time counterparts (Phelan 1990; Bulla and Muliere 2007; Zhao and Hu 2013), the literature on inferential or predictive approaches for discrete-time semi-Markov process is sparse (Barbu and Limnios, 2009, Chapter 4).

The available literature focuses on processes with a finite state space. From the frequentist perspective, Satten and Sternberg (1999) and Barbu and Limnios (2009) construct non-parametric estimators of the transition probabilities or the distributions of the holding times and study their asymptotic properties. From the Bayesian perspective, specific parametric models have been used in different settings (Patwardhan et al., 1980; Schiffman et al., 2007; Masala, 2013; Mitchell et al., 2011), but no general non-parametric approach has been developed.

In the sequel we will introduce the *semi-Markov beta-Stacy process*, a stochastic process useful for the analysis of semi-Markov models with a finite or, extending the available literature, countably infinite state space. Our perspective is both Bayesian and non-parametric because i) the Bayesian interpretation of probability is naturally suited for representing predictive uncertainty (de Finetti, 1937; Singpurwalla, 1988), and ii) non-parametric models provide a more honest assessment of posterior uncertainty than parametric models, as the formers are less tied to potentially restrictive or arbitrary parametric assumptions which may give a false sense of posterior certainty (Müller and Mitra, 2013; Hjort et al., 2010; Phadia, 2013; Ghosal and van der Vaart, 2017).

The semi-Markov beta-Stacy process is a generalization of the *beta-Stacy process* of Walker and Muliere (1997). Its law represents a prior distribution for the law of a discrete-time semi-Markov process. We will show below that this prior is conjugate with respect to i) the accumulating observations generated by a single process and ii) the finite histories of other similar (i.e. exchangeable) processes. This property makes it is particularly easy to perform inferences and predictions for a semi-Markov process.

In particular, the predictive distributions associated to the semi-Markov beta-Stacy process are available in closed form. These prescribe how to perform probabilistic predictions for the next state of a semi-Markov process given its observed history.

More precisely, we will show that these predictive distributions correspond to

the transition kernels of a *reinforced semi-Markov process*. This is a novel kind of reinforced stochastic process which can be regarded as the discrete-time analogue of the reinforced continuous-time processes of Muliere et al. (2003) and Bulla and Muliere (2007). Here, the concept of "reinforcement" coincides with that of Coppersmith and Diaconis (1986) and Pemantle (1988, 2007): a process is reinforced if, whenever it visits a state, the same becomes more likely to be visited again in the future. Thus, reinforcement corresponds to a notion of learning from the past, a central idea in the Bayesian paradigm (Muliere et al., 2000, 2003; Bulla and Muliere, 2007; Peluso et al., 2015; Arfè et al., 2018).

To gain a deeper insight into the semi-Markov beta-Stacy process, we will characterize it using a *reinforced urn process*, i.e. a random walk over a system of reinforced urns. In the prototypical reinforced urn process, whenever a random walk visits an urn, a ball is extracted from the same. After noting its color, the extracted ball is replaced in the originating urn together with an additional ball of the same color (so the extracted color is reinforced, i.e. made more likely to be extracted in future draws from the same urn). Then, the random walk jumps to another urn determined by the extracted color. Similar urn-based processes are receiving increasing attention in Statistics and Machine Learning to construct and understand nonparametric prior distributions for a wide range of stochastic models (Blackwell and MacQueen, 1973; Doksum, 1974; Mauldin et al., 1992; Walker and Muliere, 1997; Muliere et al., 2000, 2003; Bulla and Muliere, 2007; Ruggiero and Walker, 2009; Fortini and Petrone, 2012; Bacallado et al., 2013; Peluso et al., 2015; Caron et al., 2017; Arfè et al., 2018).

In more detail, below we show how a reinforced semi-Markov process can be interpreted as a particular reinforced urn process. By appealing to the representation theorems of Muliere et al. (2000) and Blackwell and MacQueen (1973), we also show the following characterization: if the future of a recurrent process (i.e. a process visiting all its states infinitely often) is predicted through the transition kernels of a reinforced semi-Markov process, then it will be as if i) the process being predicted is semi-Markov and ii) a semi-Markov beta-Stacy process prior is assigned to its probability law.

Before proceeding, we introduce some notational conventions. First, for convenience, if F is a non-decreasing function on the integers (adjoined with the σ-algebra of all subsets), then the symbol F will also be used to represent the corresponding

induced measure. Hence, for example, $F(b) - F(a) = F((a, b])$ for all $a < b$, where the interval $(a, b]$ must be interpreted as the set of all integers x such that $a < x \leq b$. Second, if $x = (x_1, x_2, \ldots)$ is a finite or infinite sequence, we denote with $x_{a:b}$ either the subsequence $(x_a, \ldots, x_b)$ of length $b - a + 1$ if $a \leq b$ or, with some abuse of notation, the empty sequence of length 0 if $a > b$. Third and last, we adopt the standard conventions so that empty sums and products are respectively equal to 0 and 1.

The reminder of the paper is structured as follows. In Section 5.2 we define discrete-time semi-Markov processes and introduce several key notations. In Section 5.3 we introduce the semi-Markov beta-Stacy process prior. In Section 5.4 we derive the corresponding posterior distributions and show that this process prior is conjugate. In Section 5.5 we introduce reinforced semi-Markov process and show that these correspond to the predictive distributions obtained from the semi-Markov beta-Stacy process prior. In Section 5.6 we characterize the semi-Markov beta-Stacy process using a system of reinforced urns. In Section 5.7, we illustrate several generalizations, each based on alternative urn constructions. In Section 5.8 we illustrate the semi-Markov beta-Stacy process prior in a simulation study. Lastly, in Section 5.9 we provide some concluding remarks and point to possible applications of our work.

5.2 Semi-Markov processes: definition and basic properties

In the sequel, let E be a non-empty finite or countably infinite set, adjoined with the discrete topology $\mathcal{E}$ of all its subsets.

Definition 5.2.1. *Let $\mathbf{P} = (P^{i,j})_{i,j \in E}$ be a transition matrix on E such that $P^{i,i} = 0$ for all $i \in E$ and let $\mathbf{F} = (F^i(\cdot) : i \in E)$ be a collection of probability distribution functions with support on the set of positive integers. Fixed l_0 in E, let the stochastic process $(L, T) = (L_n, T_n)_{n \geq 0}$ be such that $\mathbb{P}(L_0 = i | (\mathbf{P}, \mathbf{F})) = I\{i = l_0\}$ and*

$$\mathbb{P}(L_{n+1} = j, T_n \leq t | L_n = i, L_{0:n-1}, T_{0:n-1}, (\mathbf{P}, \mathbf{F})) = F^i(t) P^{i,j}$$

almost surely for all integers $n \geq 0$, $t \geq 1$, and all $i, j \in E$. Then (L, T) will be called a discrete-time Markov renewal process *starting at l_0 with* characteristic

couple $(\mathbf{P}, \mathbf{F})$. *Suppressing the dependence on* l_0*, we write* $(L, T) \sim MR(\mathbf{P}, \mathbf{F})$.

Remark 5.2.1. In Definition 5.2.1, the holding time T_k depends only on the current state L_k and not on the following state L_{k+1}. More generally, T_k may depend on both L_k and L_{k+1} (Barbu and Limnios, 2009). This can be represented by substituting the distribution F^i in Definition 5.2.1 with one of the form $F^{i,j}$ and letting $\mathbf{F} = (F^{i,j}(\cdot) : i, j \in E, i \neq j)$. Each alternative may be more or less appropriate for different applications. For simplicity, we focus on the specification of Definition 5.2.1. In Section 5.7, we will describe how to generalize our results to cover the other case.

Definition 5.2.2. *If* $(L, T) \sim MR(\mathbf{P}, \mathbf{F})$*, define* $\tau_0 = 0$ *and* $\tau_{n+1} = \sum_{h=0}^{n} T_h$ *for all* $n \geq 0$*. Then, the process* $(S_t)_{t\geq 0}$ *defined by* $S_t = L_{N(t)}$*, where* $N(t) = \sum_{n=1}^{+\infty} I\{\tau_n \leq t\}$ *for all integers* $t \geq 0$*, is the (discrete-time)* semi-Markov Process *associated to* (L, T)*,* $S = (S_t)_{t\geq 0} \sim SM(\mathbf{P}, \mathbf{F})$ *in symbols. The times* $(\tau_n)_{n\geq 1}$ *are the* jump times *of* S*.*

A semi-Markov process $(S_t)_{t\geq 0}$ describes the evolution in time of some system as it goes through different discrete states. The elements of E represent the possible states. Additionally, S_t is the state occupied at time t, $N(t)$ is the number of state changes occurred up to time t, τ_n is the time of the n-th state change, and T_k is the length of time the system spends in its k-th state (so the system first visits its $k+1$-th state at time $\tau_k + T_k$). These interpretations are possible because the assumption that $P^{i,i} = 0$ for all $i \in E$ implies that $L_k \neq L_{k+1}$ for all k with probability 1.

Example 5.2.1. *Mitchell et al. (2011) use a semi-Markov process with state-space* $E = \{$*"infected", "not infected"*$\}$ *to model the time changes in the Human Papilloma Virus status of patients who may go through several infection periods. Here* $S_t \in E$ *is the infection status of an individual at time* t*,* $N(t)$ *is the number changes in the infection status that an individual experienced by time* t*,* τ_k *is the time of the* k*-th change in the infection status of a patient,* T_k *is the length of time occurring between the* k*-th and* $k+1$*-th changes in infection status, and* $L_k \in E$ *is the infection status of a patient after this changes for the* k*-th time. For example, if at time* τ_k *the patient becomes infected (*$L_k =$ *"infected"), then* T_k *is the length of time before the patient will become infection-free again (*$L_{k+1} =$ *"not infected").*

Example 5.2.2. *Barbu and Limnios (2009, Sections 3.4) consider a semi-Markov model to describe the operation of a textile factory. To reduce pollution, the factory waste is treated in a disposal unit before being eliminated. To avoid stopping the factory, if the disposal unit fails, waste is temporarily stored in a tank. If the disposal unit is repaired before the tank is full, the factory continues operating and the tank is immediately purged. Otherwise, the whole factory must stop and a certain time is necessary to restart it. The state space of the process is thus* $E = \{1, 2, 3\}$*:* 1 *represents the state where the factory if fully operational and the tank is empty,* 2 *represents the state where the disposal unit is malfunctioning but the factory is still operational (i.e. the tank is not full), and* 3 *represents the state where the factory is stopped. Additionally, it is* $P^{1,3} = P^{3,2} = 0$*. The distribution the time until the next disposal unit failure (i.e. the holding time of the state* 1*) is* $F^1(\cdot)$*, the distribution of the time until a malfunctioning disposal unit is either restored or when it fully breaks down (i.e. the holding time of state* 2*) is* $F^2(\cdot)$*, while the distribution of the time required to restart the factory after the tank fills up (i.e. the holding time of the state* 3*) is* $F^3(\cdot)$*.*

To highlight the relation between Semi-Markov and Markov chains, suppose $S \sim SM(\mathbf{P}, \mathbf{F})$ is such that $F^i(\{t\}) = p_i(1-p_i)^{t-1}$ for all integers $t \geq 1$ and some $p_i \in (0, 1)$ (i.e. the holding times of the state $i \in E$ are geometrical distributed with parameter p_i). Then S is a (homogeneous) Markov chain such that $\mathbb{P}(S_{t+1} = j | S_{0:t-1}, S_t = i) = p_i P^{i,j}$ for all $j \in E$, $j \neq i$ and $\mathbb{P}(S_{t+1} = i | S_{0:t-1}, S_t = i) = 1 - p_i$ for all $t \geq 1$. Conversely, if S is a Markov chain with transition matrix $(p_{i,j})_{i,j \in E}$, then $S \sim SM(\mathbf{P}, \mathbf{F})$ with $P^{i,j} = p_{i,j}/(1-p_{i,i})$ for all $j \neq i$, $P^{i,i} = 0$, and $F^i(t) = (1-p_{i,i})p_{i,i}^{t-1}$ for all $t \geq 1$.

Note that, since $P^{i,i} = 0$ and F^i has support on the positive integers for all $i \in E$, the semi-Markov process S cannot have absorbing states, i.e. states such that $S_t = i$ for all sufficiently large $t \geq 0$ with positive probability. This assumption simplifies our analysis, although it might be restrictive for some applications. The presence of an absorbing state i could be allowed by letting $P^{i,i} = 1$ and $F^i(\{+\infty\}) = 1$. With additional effort, the results in the following sections could be extended to this case as well.

Remark 5.2.2. Knowing $S_{0:t}$ is equivalent to knowing the values of $N(t)$, $L_{0:N(t)}$, $\tau_{1:N(t)}$, and that $\tau_{N(t)+1} > t$. Furthermore, denote $l(t) = t - \tau_{N(t)} = \max\{k =$

$0, 1, \ldots, t : S_t = S_{t-1} = \cdots = S_{t-k}\}$ the time spent by S in the state S_t just prior to time t. Then knowing $S_{0:t}$ is the same as knowing the values of $N(t)$, $L_{0:N(t)}$, $T_{0:N(t)-1}$, and that $T_{N(t)} > l(t)$.

Example 5.2.3. *To exemplify, note that observing* $S_{0:5} = (i_0, i_0, i_1, i_2, i_2, i_2)$ *for some distinct* $i_0, i_1, i_2 \in E$ *is equivalent to observing* $N(5) = 2$, $L_0 = i_0$, $L_1 = i_1$, $L_2 = i_2$, $\tau_1 = 2$, $\tau_2 = 3$, $\tau_3 > 5$, $l(5) = 2$, $T_0 = 2$, $T_1 = 1$, *and* $T_2 \geq 3$, *i.e.* $T_2 > 2 = l(5)$.

5.3 The semi-Markov beta-Stacy prior

From a Bayesian nonparametric perspective (Ferguson, 1973; Hjort et al., 2010; Müller and Mitra, 2013), a prior distribution on the law of a semi-Markov process $S \sim SM(\mathbf{P}, \mathbf{F})$ is the law of a stochastic process whose sample paths are characteristic couples $(\mathbf{P}, \mathbf{F})$ with probability 1. The semi-Markov beta-Stacy process is one such stochastic process. Our strategy to define it is to separately assign a nonparametric prior distribution to i) each holding time distribution F^i and ii) the transition matrix $\mathbf{P}$.

As a starting point, we consider the discrete-time beta-Stacy process of Walker and Muliere (1997), a common Bayesian nonparametric prior for time-to-event distributions (Singpurwalla, 2006; Bulla and Muliere, 2007; Rigat and Muliere, 2012; Arfè et al., 2018). The beta-Stacy process will be used as the prior for the holding time distributions F^i.

Definition 5.3.1 (Walker and Muliere (1997))**.** *Let* $c(t)$ *be a positive real number for all integer* $t > 0$. *Also let* F_0 *be a probability distribution function with support on the set of positive integers. A random cumulative distribution function* F *with support on the set of positive integers is said to be a* beta-Stacy process $BS(c, F_0)$ *if there exists a sequence* $(U_t)_{t\geq 1}$ *of independent random variables such that i) for all integers* $t \geq 1$,

$$U_t \sim Beta(c(t)F_0(\{t\}), c(t)F_0((t, +\infty)));$$

ii) $F((t, +\infty)) = \prod_{k=1}^{t}(1 - U_k)$ *for all integers* $t \geq 0$.

Remark 5.3.1. If $F \sim BS(c, F_0)$, then $\mathbb{E}[F(t)] = F_0(t)$ and $\mathrm{Var}(F(t))$ is a decreasing function of $c(t)$ such that $\mathrm{Var}(F(t)) \to 0$ as $c(t) \to +\infty$. Hence F_0 is the mean of the process, while c controls its dispersion (Walker and Muliere, 1997).

The beta-Stacy process is especially useful thanks to its conjugacy property, which implies that the posterior distribution of $F \sim BS(c, F_0)$ conditional on a sample of exact observations from F is again a beta-Stacy process. The beta-Stacy process is also conjugate with respect to an observation which has been *censored*, i.e. whose value is only known to exceed some known constant (Kalbfleisch and Prentice, 2002; Singpurwalla, 2006). These properties are summarized in the following Proposition, which is a specific case of the more general Theorem 1 of Walker and Muliere (1997).

Proposition 5.3.1 (Walker and Muliere (1997)). If, conditionally on $F \sim BS(c, F_0)$, $T_1, \ldots, T_n$ are independently distributed according to F, then the posterior distribution of F given $T_1, \ldots, T_n$ is $BS(c_*, F_*)$, where

$$F_*((t, +\infty)) = \prod_{s=1}^{t} \left[1 - \frac{c(s)F_0(\{s\}) + N(\{s\})}{c(s)F_0([s, +\infty)) + N([s, +\infty))} \right]$$
$$c_*(t) = \frac{c(t)F_0((t, +\infty)) + N((t, +\infty))}{F_*((t, +\infty))}.$$

where $N(t) = \sum_{i=1}^{n} I\{T_i \leq t\}$. Instead, the posterior distributions of F given $T_n > t^*$ (i.e. a censored observation), where t^* is a fixed constant, is $BS(c_*, F_*)$, where now

$$F_*((t, +\infty)) = \prod_{s=1}^{t} \left[1 - \frac{c(s)F_0(\{s\})}{c(s)F_0([s, +\infty)) + I\{t^* \geq s\}} \right],$$
$$c_*(t) = \frac{c(t)F_0((t, +\infty)) + I\{t^* \geq s\}}{F_*((t, +\infty))}.$$

To specify a prior on the transition matrix $\mathbf{P}$ we will take advantage of the Dirichlet process of Ferguson (1973), a fundamental non-parametric process prior for probability measures (Hjort et al., 2010). Since the i-the row $P^i = (P^{i,j})_{j \in E}$ of $\mathbf{P}$ is the probability measure $P^i(\cdot)$ on $(E, \mathcal{E})$ defined by $P^i(\{j\}) = P^{i,j}$ for all $j \in E$, this can be assigned a Dirichlet process prior.

Definition 5.3.2 (Ferguson (1973))**.** *Let m be a measure on $(E, \mathcal{E})$ such that $0 < m(E) < +\infty$. A random probability measure P on $(E, \mathcal{E})$ is a* Dirichlet process *with* base measure *m, or $P \sim Dir(m)$ in symbols, if for every partition $A_1, \ldots, A_n \in \mathcal{E}$ of E it holds that*

$$(P(A_1), \ldots, P(A_n)) \sim Dirichlet(m(A_1)/m(E), \ldots, m(A_n)/m(E)).$$

Remark 5.3.2. If $P \sim Dir(m)$, then $P(A) \sim Beta(m(A), m(A^c))$ for all A. In particular, $\mathbb{E}[P(A)] = m(A)/m(E)$ and $\text{Var}(P(A)) \to 0$ as $m(E) \to +\infty$, so $m(E)$ controls the dispersion of $P(\cdot)$ around its mean $m(\cdot)/m(E)$. Additionally, if $A \in \mathcal{E}$ is such that $m(A) = 0$, then $P(A) \sim Beta(0, m(E))$ and so $P(A) = 0$ almost surely.

The Dirichlet process is a particular case of the beta-Stacy process. In particular, since in our setting E is countable, this can be identified with a set of the form $E = \{1, 2, \ldots, k\}$ for some $k \leq +\infty$. With this identification, let m be a measure on $(E, \mathcal{E})$ such that $0 < m(E) < +\infty$ and let $P \sim Dir(m)$. The probability measure P is entirely determined by its distribution function $F(t) = \sum_{x \in E:\ x \leq t} P(\{x\})$. Following the same reasoning as in Walker and Muliere (1997, Remark 5), it can be shown that $F \sim BS(c, F_0)$, where $c(t) = m(E)$ for all integers $t > 0$ and F_0 is determined by $F_0(\{t\}) = m(\{t\})/m(E)$ for all integers t such that $0 < t \leq k$, and $F_0(\{t\}) = 0$ for $t > k$.

Akin as the beta-Stacy process, the Dirichlet process is also conjugate, as highlighted by the following proposition. This could be proved either by representing the Dirichlet process as a specific case of the beta-Stacy process, or by appealing to Theorem 1 of Ferguson (1973) and the facts that E is countable and $\mathcal{E}$ is its power set.

Proposition 5.3.2 (Ferguson (1973)). Suppose that $P \sim Dir(m)$ and, conditionally on P, $X_1, \ldots, X_n$ are independently distributed with common law P. Then the posterior distribution of P given $X_1, \ldots, X_n$ is $Dir(m_*)$, where m_* is the measure on E determined by

$$m_*(\{i\}) = m(\{i\}) + \sum_{j=1}^{n} I\{X_j = i\}$$

for all $i \in E$.

Having introduced all required elements, we are finally ready to define the semi-Markov beta-Stacy process. To do so, let $m^i(\cdot)$ be a measure on $(E, \mathcal{E})$ such that $0 < m^i(E) < +\infty$ and $m^i(\{i\}) = 0$ for all $i \in E$. Let $c^i(t)$ be a positive real number for any integer $t > 0$. Also let F_0^i be a distribution function with support on the set of positive integers for all $i \in E$. Lastly, let $\mathbf{m} = (m^i)_{i \in E}$, $\mathbf{c} = (c^i)_{i \in E}$, and $\mathbf{F_0} = (F_0^i)_{i \in E}$.

Definition 5.3.3. *A random characteristic couple* $(\mathbf{P}, \mathbf{F})$ *has a* semi-Markov beta-Stacy *distribution with parameters* $(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$*, or* $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$*, if:*

1. $\mathbf{P}$ *and* $\mathbf{F}$ *are independent;*

2. *the rows* $P^i(\cdot)$, $i \in E$, *of* $\mathbf{P}$ *are independent;*

3. *the distributions* F^i, $i \in E$, *in* $\mathbf{F}$ *are independent;*

4. $P^i(\cdot)$ *is a Dirichlet process with base measure* m^i *for all* $i \in E$: $P^i(\cdot) \sim Dir(m^i)$;

5. *for all* $i \in E$, F^i *is a beta-Stacy process with precision parameters* c^i *and centering distribution* F_0^i: $F^i \sim BS(c^i, F_0^i)$.

Note that each realization of $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ is a valid characteristic couple, justifying the use of the law of a semi-Markov beta-Stacy process as a prior distribution for a characteristic couple $(\mathbf{P}, \mathbf{F})$.

More precisely, if $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$, then with probability 1, i) $F^i(\cdot)$ is a cumulative distribution function with support on the positive integers and ii) $\mathbf{P}$ is a transition matrix on E such that $P^{i,i} = 0$ for all $i \in E$. The first point follows directly from the properties of the beta-Stacy process. The second point instead follows because each realization of the Dirichlet process is almost surely a probability measure. This implies that $0 \leq P^{i,j} = P^i(\{j\}) \leq 1$ and $\sum_{j \in E} P^{i,j} = \sum_{j \in E} P^i(\{j\}) = P^i(E) = 1$ for all $i, j \in E$ with probability 1. Since $m^i(\{i\}) = 0$ for all $i \in E$, it must also be $P^{i,i} = 0$ almost surely by Remark 5.3.2.

More generally, it will be $P^i(\{j\}) = P^{i,j} = 0$ almost surely for all $j \in E$ such that $m^i(\{j\}) = 0$. In this case, each realization of a $SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ will be the law of a semi-Markov process which cannot perform transition from i to j.

5.4 Posterior computations

We will now prove that the semi-Markov beta-Stacy process prior is conjugate. To do so, we will need to introduce some additional notions.

Consider a finite sequence of states $s_{0:t} = (s_0, \ldots, s_t) \in E^{t+1}$. For each $i \in E$, any maximal subsequences $s_{a:b}$ $(0 \leq a \leq b \leq t)$ such that $s_c = i$ for all $a \leq c \leq b$ will be called an i-block of $s_{0:t}$. In particular, an i-block $s_{a:b}$ will be called *terminal* if $b = t$, *non-terminal* otherwise. Suppose now that $S \sim SM(\mathbf{P}, \mathbf{F})$. Moreover, $N^{i,t}(l)$ will denote the number of non-terminal i-blocks of length $\leq l$ present in $S_{0:t} \in E^{t+1}$.

Additionally, for all $i, j \in E$, $i \neq j$, let $M^{i,j}(t) = \sum_{k=1}^{t} I\{S_{k-1} = i, S_k = j\}$ be the number of transitions from state i to state j in $S_{0:t}$.

Remark 5.4.1. As observed in Remark 5.2.2, knowing $S_{0:t}$ is equivalent to knowing the values of $N(t)$, $L_{0:N(t)}$, $\tau_{1:N(t)}$, $T_{0:N(t)-1}$, and that $T_{N(t)} > l(t)$. This implies that the terminal block of $S_{0:t}$ is an $L_{N(t)}$-block $S_{t-l(t):t}$ of length $x(t)$. Additionally, $S_{0:t}$ contains exactly $N(t)$ non-terminal blocks. For $k = 0, \ldots, N(t) - 1$, the $k+1$-th of such non-terminal blocks is the L_k-block $S_{\tau_k:\tau_{k+1}}$ of length T_k. Consequently,

$$N^{i,t}(l) = \sum_{k=0}^{N(t)-1} I\{T_k \leq l, L_k = i\}.$$

Example 5.4.1. *Going back to Example 5.2.3, the blocks of* $S_{0:5} = (i_0, i_0, i_1, i_2, i_2, i_2)$ *are the non-terminal* i_0*-block* $S_{0:1}$ *of length* $T_0 = 2$*, the non-terminal* i_1*-block* $S_{2:2}$ *of length* $T_1 = 1$*, and the terminal* i_2*-block* $S_{3:5}$ *of length* $3 = l(5) + 1$*. Additionally:* $N^{i_0,5}(1) = 0$*,* $N^{i_0,5}(2) = N^{i_0,5}(l) = 1$ *for all* $l \geq 2$*;* $N^{i_1,5}(1) = N^{i_1,5}(l) = 1$ *for all* $l \geq 1$*;* $N^{i,5}(l) = 0$ *for all* $l > 0$ *if* $i \neq i_0, i_1$*; and* $M^{i_0,i_1}(5) = M^{i_1,i_2}(5) = 1$.

With these notations, we can now state the following theorem.

Theorem 5.4.1. *Suppose that, given* $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$*, it is* $S \sim SM(\mathbf{P}, \mathbf{F})$*. Then, the posterior distribution of* $(\mathbf{P}, \mathbf{F})$ *given* $S_{0:t} = i_{0:t}$ *is* $SMBS(\mathbf{m}_*, \mathbf{c}_*, \mathbf{F}_*)$*, where:*

1. *For all* $i \in E$*,* m^i_* *is defined by* $m^i_*(\{j\}) = m^i(\{j\}) + M^{i,j}(t)$ *for* $j \in E$*,* $j \neq i$.
2. *For all* $i \in E$*,* $i \neq i_t$*,* F^i_* *and* c^i_* *are determined by letting*

$$F^i_*((u, +\infty)) = \prod_{s=1}^{u} \left[1 - \frac{c^i(s)F^i_0(\{s\}) + N^{i,t}(\{s\})}{c^i(s)F^i_0([s, +\infty)) + N^{i,t}([s, +\infty))}\right]$$

$$c^i_*(u) = \frac{c^i(u)F^i_0((u, +\infty)) + N^{i,t}((u, +\infty))}{F_*((u, +\infty))}$$

   for each integer $u > 0$.

3. *For* $i = i_t$*,* F^i_* *and* c^i_* *are instead determined by letting*

$$F^i_*((u, +\infty)) = \prod_{s=1}^{u} \left[1 - \frac{c^i(s)F^i_0(\{s\}) + N^{i,t}(\{s\})}{c^i(s)F^i_0([s, +\infty)) + N^{i,t}([s, +\infty)) + I\{l(t) \geq s\}}\right]$$

$$c^i_*(u) = \frac{c^i(u)F^i_0((u, +\infty)) + N^{i,t}((u, +\infty)) + I\{l(t) \geq u\}}{F_*((u, +\infty))}$$

for each integer $u > 0$.

Proof. To begin, note that by Remark 5.2.2 observing a sequence of states $S_{0:t} = i_{0:t}$ such that $N(t) = n$ is equivalent to observing $L_{0:n} = l_{0:n}$, $T_{0:n-1} = t_{0:n-1}$, and $T_n > l(t)$, where $l_{0:n}$ is the sequence of distinct states in $i_{0:t}$ (in the same order) and the times $t_{0:n-1}$ are determined uniquely by the position of the the state changes in the sequence $i_{0:t}$. Consequently, by Remark 5.4.1 the likelihood function associated to the observation of $S_{0:t} = i_{0:t}$ is given by

$$\begin{aligned}
\mathbb{P}\left(S_{0:t} = i_{0:t} | \mathbf{P}, \mathbf{F}\right) &= \mathbb{P}\left(L_{0:n} = l_{0:n}, T_{0:n-1} = t_{0:n-1}, T_n > l(t) | \mathbf{P}, \mathbf{F}\right) \\
&= F^{l_0}(t_0) \left[\prod_{k=1}^{n-1} F^{l_k}(\{t_k\}) P^{l_{k-1}, l_k}\right] \cdot \left[F^{l_n}((l(t), +\infty))\right] \\
&= \left[\prod_{i \in E} \prod_{s=1}^{t} F^i(\{s\})^{N^{i,t}(\{s\})} F^i((l(t), +\infty))^{I\{i = l_n\}}\right] \cdot \\
&\qquad \left[\prod_{\substack{i,j \in E \\ i \neq j}} P^i(\{j\})^{M^{i,j}(t)}\right]
\end{aligned}$$

Since the likelihood can be factorized as the product of individual terms depending only on F^i or $P^i(\cdot)$ for some i, by points 1-3 of Definition 5.3.3 it follows that, conditionally on $S_{0:t} = i_{0:t}$, i) $\mathbf{P}$ and $\mathbf{F}$ are independent, ii) the rows $P^i(\cdot)$, $i \in E$, of $\mathbf{P}$ are independent, and iii) the distributions F^i, $i \in E$, in $\mathbf{F}$ are independent.

It can now be seen that: i) the posterior distribution of F^i, $i \neq l_n$ depends only on the observed values of those T_k such that $L_k = i$ and it is the same as if these value were obtained as a random sample of independent and identically distributed observations from F^i; ii) the same is true for the posterior distribution of F^{l_n} except that T_n is censored, as only $T_n > l(t)$ is known; iii) the posterior distribution of $P^i(\cdot)$ depends only on each and only those l_k in the sequence $l_{0:n}$ which are preceded by the state i; the corresponding posterior distribution is the same as if these were a random sample from $P^i(\cdot)$. The thesis now follows from Propositions 5.3.1 and 5.3.2. □

Theorem 5.4.1 allows to compute the posterior distribution of $(\mathbf{P}, \mathbf{F})$ associated to the observation of the history $S_{0:t}$ up to time t of some semi-Markov process

$S \sim SM(\mathbf{P}, \mathbf{F})$. For example, in the context of Example 5.2.2, the history $S_{0:t}$ may represent the (unreplicable) history of failures in the operation of the textile factory.

In some settings, however, multiple independent semi-Markov processes S^1, ..., $S^n \sim SM(\mathbf{P}, \mathbf{F})$ may be observed up to fixed time points $t^1, \ldots, t^n$, generating data $S^1_{0:t^1}$, ..., $S^n_{0:t^n}$. For instance, in the context of Example 5.2.1, $S^1_{0:t^1}$, ..., $S^n_{0:t^n}$ may represent the histories of infection status of n independent patients. In this case, the posterior distribution of $(\mathbf{P}, \mathbf{F})$ is obtained by iteratively applying Theorem 5.4.1.

It could also be shown that Theorem 5.4.1 remains valid also if the the process S is observed up to some stopping time τ, so that the posterior distribution of $(\mathbf{P}, \mathbf{F})$ given $S_{0:\tau}$ is the semi-Markov beta-Stacy process obtained by applying Theorem 5.4.1 after substituting $S_{0:\tau}$ for $S_{0:t}$. Following an argument similar as those presented by Heitjan and Rubin (1991), the same result also holds if τ is a random variable a priori independent of S and $(\mathbf{P}, \mathbf{F})$.

5.5 Predictive distributions and reinforced semi-Markov processes

We now address the problem of predicting the evolution of a process $S \sim SM(\mathbf{P}, \mathbf{F})$. Specifically, assuming $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$, we derive the one-step-ahead predictive distributions of S, i.e. the conditional distributions $\mathbb{P}(S_{t+1} = \cdot | S_{0:t})$ for $t \geq 0$. These play an important role in applications. For instance, in Example 5.2.1 they allow to predict the future infection status of an individual patient given its history of infections. Instead, in Example 5.2.2, they allow to quantify the future risk that the textile factory will have to stop its operations.

Theorem 5.5.1. *Suppose that, given* $(\mathbf{P}, \mathbf{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$*, it is* $S \sim SM(\mathbf{P}, \mathbf{F})$*. Define for simplicity* $x(t) = l(t) + 1$ *for all integers* $t \geq 0$*. Then, with probability 1,* $\mathbb{P}(S_{t+1} = \cdot | S_{0:t}) = k_t(S_{0:t}; \cdot)$ *for all integers* $t \geq 0$*, where, letting* $S_t = i$*,* k_t *is the*

transition kernel defined as follows:

$$k_t(S_{0:t};i) = \frac{c^i(x(t))F_0^i((x(t),+\infty)) + N^{i,t}((x(t),+\infty))}{c^i(x(t))F_0^i([x(t),+\infty)) + N^{i,t}([x(t),+\infty))}$$
$$k_t(S_{0:t};j) = \frac{c^i(x(t))F_0^i(\{x(t)\}) + N^{i,t}(\{x(t)\})}{c^i(x(t))F_0^i([x(t),+\infty)) + N^{i,t}([x(t),+\infty))} \cdot \frac{m^i(\{j\}) + M^{i,j}(t)}{m^i(E) + \sum_{h\neq i} M^{i,h}(t)},$$

for all $j \neq i$.

Proof. Suppose that, conditionally on $(\mathbf{P},\mathbf{F}) \sim SMBS(\mathbf{m},\mathbf{c},\mathbf{F_0})$, $S \sim SM(\mathbf{P},\mathbf{F})$. To prove the thesis, observe that by Remark 5.2.2 it is

$$\mathbb{P}\left(S_{t+1}=j|S_{0:t},\mathbf{P},\mathbf{F}\right) = \mathbb{P}\left(L_{N(t+1)}=j|N(t),L_{0:N(t)},T_{0:N(t)-1},T_{N(t)}>l(t),\mathbf{P},\mathbf{F}\right)$$

almost surely. Consequently, on the event $\{N(t)=n, L_{0:n}=i_{0:n}, T_{0:n-1}=t_{0:n-1}\}$ with $j=i_n$ it is

$$\begin{aligned}\mathbb{P}\left(S_{t+1}=j|S_{0:t},\mathbf{P},\mathbf{F}\right) &= \\ \mathbb{P}\left(T_n>x(t)|L_{0:n}=i_{0:n},T_{0:n-1}=t_{0:n-1},T_n>l(t),\mathbf{P},\mathbf{F}\right) &= \\ = \frac{F^{i_n}((x(t),+\infty))}{F^{i_n}((l(t),+\infty))}.&\end{aligned}$$

Since F^{i_n} has a beta-Stacy distribution, by Theorem 5.4.1 conditionally on $S_{0:t}$ it is

$$F^{i_n}((x(t),+\infty)) = \prod_{k=1}^{x(t)}(1-U_k) = (1-U_{x(t)})F^{i_n}((l(t),+\infty))$$

for independent $U_1,\ldots,U_{x(t)}$ such that

$$U_{x(t)} \sim Beta\left(c_*^{i_n}(x(t))F_*^{i_n}(\{x(t)\}), c_*^{i_n}(x(t))F_*^{i_n}((x(t),+\infty))\right).$$

Thus, from Theorem 5.4.1,

$$\begin{aligned}
\mathbb{P}\left(S_{t+1}=j|S_{0:t}\right) &= \mathbb{E}\left[\mathbb{P}\left(S_{t+1}=j|S_{0:t},\mathbf{P},\mathbf{F}\right)|S_{0:t}\right] \\
&= \mathbb{E}\left[1-U_{x(t)}|S_{0:t}\right] \\
&= \frac{F_*^{i_n}((x(t),+\infty))}{F_*^{i_n}([x(t),+\infty))} \\
&= 1-\frac{c^{i_n}(x(t))F_0^{i_n}(\{x(t)\})+N^{i_n,t}(\{x(t)\})}{c^{i_n}(x(t))F_0^{i_n}([x(t),+\infty))+N^{i_n,t}([x(t),+\infty))+I\left\{l(t)\geq x(t)\right\}} \\
&= k_t(S_{0:t},j)
\end{aligned}$$

as needed.

Continuing, on the event $\{N(t)=n, L_{0:n}=i_{0:n}, T_{0:n-1}=t_{0:n-1}\}$ with $j\neq i_n$, $\mathbb{P}\left(S_{t+1}=j|S_{0:t},\mathbf{P},\mathbf{F}\right)$ equals

$$\begin{aligned}
&\mathbb{P}\left(T_n=x(t), L_{n+1}=j|L_{0:n}=i_{0:n}, T_{0:n-1}=t_{0:n-1}, T_n>l(t),\mathbf{P},\mathbf{F}\right)= \\
&\quad= \frac{F^{i_n}(\{x(t)\})}{F^{i_n}([x(t),+\infty))}\cdot P^{i_n,j} \\
&\quad= \left(1-\frac{F^{i_n}((x(t),+\infty))}{F^{i_n}([x(t),+\infty))}\right)\cdot P^{i_n,j} \\
&\quad= U_{x(t)}P^{i_n,j}.
\end{aligned} \tag{5.1}$$

By Theorem 5.4.1, $U_{x(t)}$ and $P^{i_n,j}\sim Beta(m_*^{i_n}(\{j\}), m_*^{i_n}(E\backslash\{j\}))$ are independent given on $S_{0;t}$. The thesis now follows by taking expectations conditionally on $S_{0:t}$. □

By the Ionescu-Tulcea Theorem (Çinlar, 2011, Theorem 4.7), the sequence of predictive distributions k_t defines the law of a new stochastic process:

Definition 5.5.1. *A stochastic process $S=(S_t)_{t\geq 0}$ with state space $(E,\mathcal{E})$ is called a reinforced semi-Markov process with parameters $(\mathbf{m},\mathbf{c},\mathbf{F_0})$, or $S\sim RSM(\mathbf{m},\mathbf{c},\mathbf{F_0})$, if $\mathbb{P}\left(S_0=l_0\right)=1$ and $\mathbb{P}\left(S_{t+1}=j|S_{0:t}\right)=k_t(S_{0:t};j)$ almost surely for all $j\in E$ and $t\geq 0$.*

With this definition, the following is a trivial corollary of Theorem 5.5.1:

Corollary 5.5.1. *If, conditionally on $(\mathbf{P},\mathbf{F})\sim SMBS(\mathbf{m},\mathbf{c},\mathbf{F_0})$, $S\sim SM(\mathbf{P},\mathbf{F})$, then marginally it is $S\sim RSM(\mathbf{m},\mathbf{c},\mathbf{F_0})$.*

Compatibly with the definition of Coppersmith and Diaconis (1986) and Pemantle (1988, 2007), the process $S \sim RSM(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ is "reinforced" in the following sense: if S performs a transition from a state i to a state $j \neq i$, this becomes more likely in the future. More precisely, say that $S_t = i$ and consider the probability $k_t(S_{0:t}, j)$ for some $j \neq i$. By Equations (5.1), $k_t(S_{0:t}, j)$ is increasing in $M^{i,j}(t)$, i.e. the number of times that a transition from i to j already occurred by time t.

5.6 Predictive characterization by reinforced urn processes

A sequence $(L_n)_n \geq 1$ of random elements of E is said to be a *Pòlya sequence* generated by a measure $m(\cdot)$ on E if it is the result of successive draws from a *generalized Pòlya urn* whose initial composition is determined by m.

Specifically, this is a reinforced urn U which initially contains $m(\{i\})$ balls of color $i \in E$. Balls are repeatedly extracted from the urn and, after every draw, the extracted ball is replaced together with another additional ball of the same color. The color of the ball extracted at the n-th draw gives the value of L_n, so $\mathbb{P}(L_1 = i) = m(\{i\})/m(E)$ and, for all $n \geq 1$,

$$\mathbb{P}(L_{n+1} = i | L_{1:n}) = \frac{m(\{i\}) + \sum_{h=1}^{n} I\{L_h = i\}}{m(E) + n}.$$

The seminal results of Blackwell and MacQueen (1973) imply that $(L_n)_{n \geq 1}$ is exchangeable and its *de Finetti measure* is $Dir(m)$. In other words, there exists a random probability measure $P \sim Dir(m)$ such that the L_n are independent and have common distribution $P(\cdot)$, conditionally on $P(\cdot)$.

Spurring from the work of Blackwell and MacQueen (1973), other urn models have been used to characterize many other common nonparametric prior processes. For example, using models based on Pòlya urns it is possible to generate Pòlya trees (Mauldin et al., 1992) or the beta-Stacy process (Walker and Muliere, 1997). Fortini and Petrone (2012) provide references to other modern examples. Many of these constructions can be unified using the reinforced urn processes of Muliere et al. (2000), which also provide a tool to characterize general neutral-to-the-right processes (Doksum, 1974).

Of particular interest to us is the following urn scheme characterizing the discrete-time beta-Stacy process. Here, let $c(t)$ be a positive real number for all integer $t > 0$ and F_0 be a distribution function with support on the positive integers.

Suppose V_1, V_2, V_3, ..., V_k, ... is an infinite sequence of Pòlya urns. Each urn V_k contains $c(t)F_0(\{t\})$ black balls and $c(t)F_0((t, +\infty))$ white balls. As before, every time a ball is extracted from an urn, it is replaced together with another ball of the same color.

Starting from V_1, for $k \geq 1$ sample a ball from V_k. If its color is white, continue sampling from V_{k+1}, otherwise set $T_1 = k$ and return to V_1 after having reinforced all visited urns. Restarting from V_1 and repeating the process it is possible to generate the variables T_2, T_3, T_4, and so on. It is possible to show that the urn V_1 is *recurrent*, i.e. it is visited infinitely often with probability 1. Consequently, this scheme generates an infinite sequence $(T_n)_{n\geq 1}$ of random variables such that

$$\mathbb{P}(T_{n+1} > t | T_{1:n}) = \prod_{s=1}^{t} \left[1 - \frac{c(s)F_0(\{s\}) + N(\{s\})}{c(s)F_0([s, +\infty)) + N([s, +\infty))} \right],$$

where $N(t) = \sum_{i=1}^{n} I\{T_i \leq t\}$ (the right-hand side is exactly F_* from Proposition 5.3.1).

Here, Muliere et al. (2000) have shown that $(T_n)_{n\geq 1}$ is exchangeable and its de Finetti measure is the $BS(c, F_0)$ distribution. Hence, there exists a random $F(\cdot) \sim BS(c, F_0)$ such that the T_n are independent and have distribution $F(\cdot)$, conditional on $F(\cdot)$.

Definition 5.6.1. *For simplicity, we will say that a generalized Pòlya urn U like the one used above to characterize the $Dir(m)$ process is a* $Dir(m)$-urn. *Similarly, we say that a system V of reinforced urns V_1, V_2, V_3, ... like the one used to characterize the $BS(c, F_0)$ process is a* $BS(c, F_0)$-system.

We can now describe an urn-based characterization of the semi-Markov beta-Stacy process. To do so, associate every $i \in E$ with a $Dir(m^i)$-urn U_i and a $BS(c^i, F_0^i)$-system V_i made up of the urns $V_{i,1}$, $V_{i,2}$, $V_{i,3}$, and so on. Generate a sequence $\{(L_k, T_k)\}_{k\geq 0}$ as follows. Set $L_0 = l_0$. Then, for all $k \geq 0$, generate T_k from V_{L_k} as above, and, independently, set L_{k+1} to the color of the ball extracted from U_{L_k}. This generative process is illustrated graphically in Figure 5.1

Continuing, define a process $S = (S_t)_{t\geq 0}$ with state space E as follows. Define $\tau_0 = 0$, $\tau_{n+1} = \sum_{h=0}^{n} T_h$ for all $n \geq 0$, and $N(t) = \sum_{n=1}^{+\infty} I\{\tau_n \leq t\}$ for all integers

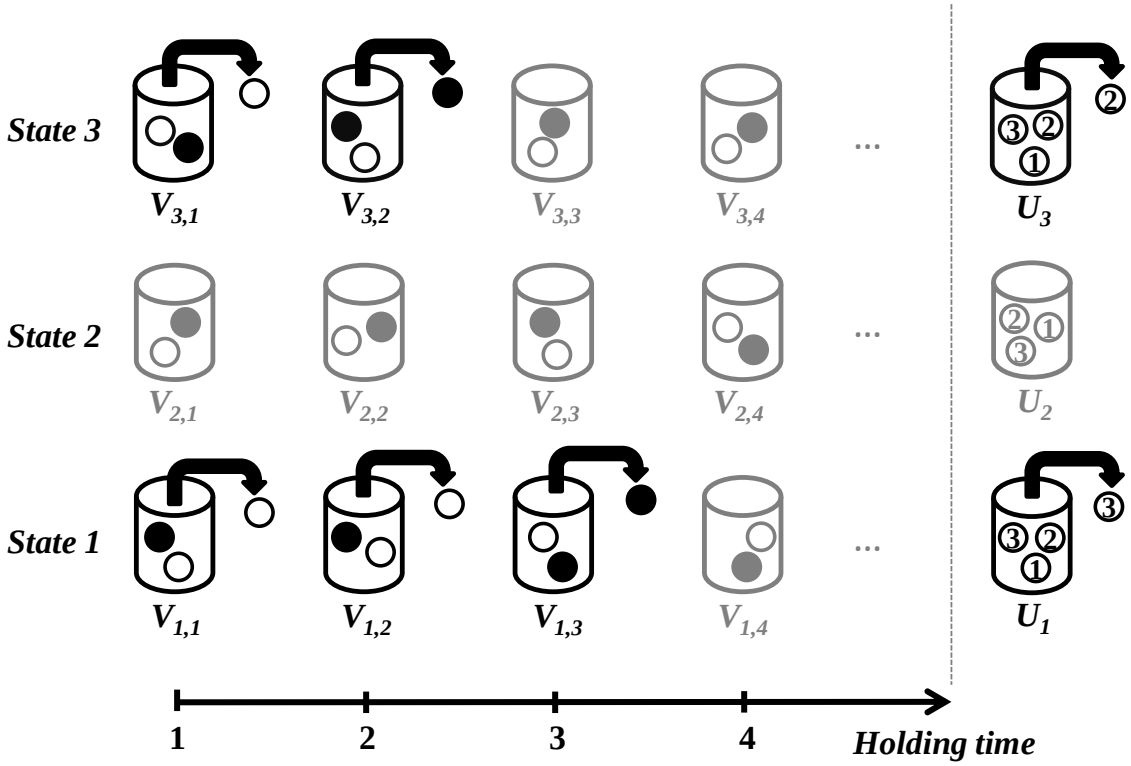


Figure 5.1: Graphical illustration of the reinforced urn process of Section 5.6. In the figure, the path of the process corresponds to the observation of $(L_0, T_0) = (1, 3)$, $(L_1, T_1) = (3, 2)$, and $L_2 = 2$. Specifically, the process starts from the urn corresponding to the value $T_0 = 1$ for the holding time of the state $L_0 = 1$. The $BS(c^1, F_0^1)$-system V_{11}, V_{12}, V_{13}, ... is traversed left to right until a black ball is extracted from V_{13}, determining the value $T_0 = 3$. The process then jumps to the $Dir(m^1)$-urn U_1, from which a ball of color "3" is extracted. Thus, $L_1 = 3$ and the process jumps to V_{31}, the first urn of the $BS(c^3, F_0^3)$-system represented in the third row of the graph. The process then resumes similarly to generate the values $T_1 = 2$ and $L_2 = 2$.

$t \geq 0$. Lastly, define the process $S = (S_t)_{t\geq 0}$ by letting $S_t = L_{N(t)}$ for all integers $t \geq 0$. It is not hard to show that $\mathbb{P}(S_{t+1} = \cdot | S_{0:t}) = k_t(S_{0:t}, \cdot)$, where the kernel k_t is the same as in Theorem 5.5.1. This shows that $S \sim RSM(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$. Clearly, any $RSM(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ process can be generated in this way.

Now, for all $i \in E$ let $v_{i,0} = -1$ and, for all integers $n \geq 1$, let $v_{i,n} = \inf\{k > v_{i,n-1} : L_k = i\}$ be the time of the n-th visit of the sequence $(L_k)_{k\geq 0}$ to the state i. The process $S = (S_t)_{t\geq 0}$ just introduced will be said to be *recurrent* if

$$\mathbb{P}\left(\bigcap_{i\in E}\bigcap_{n=1}^{+\infty}\{v_{i,n} < +\infty\}\right) = 1. \tag{5.2}$$

In other words, S is recurrent if it visits every state in E an infinite number of times with probability 1. If S is recurrent, for each $i \in E$ we can define the infinite sequence $\{(L_{i,n}, T_{i,n}) = (L_{v_{i,n}+1}, T_{v_{i,n}})\}_{n\geq 1}$. Note that $T_{i,n}$ is the (finite) length of the n-th i-block in S, which is immediately followed by a $L_{i,n}$-block. In other words, $T_{i,n}$ the length of time S stays in i during the n-th visit to that state, while $L_{i,n}$ is the state visited by S immediately after its n-th visit to i is over.

With these notions, we can now show the following partial converse of Corollary 5.5.1:

Theorem 5.6.1. *Suppose $S \sim RSM(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ is recurrent. Then there exists a random characteristic couple $(\mathbf{P}, \boldsymbol{F})$ such that:*

1. *conditional on $(\mathbf{P}, \boldsymbol{F})$, $S \sim SM(\mathbf{P}, \boldsymbol{F})$;*
2. *$(\mathbf{P}, \boldsymbol{F}) \sim SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$.*

To show this result we will make use of the following lemma:

Lemma 5.6.1. *Suppose $S \sim RSM(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ is recurrent. Then:*

1. *the sequences $\{(L_{i,n}, T_{i,n})\}_{n\geq 1}$ for $i \in E$ are independent;*
2. *the sequences $(L_{i,n})_{n\geq 1}$ and $(T_{i,n})_{n\geq 1}$ are independent for all $i \in E$;*
3. *there exists a random probability measure $P^i \sim Dir(m^i)$ such that the $L_{i,n}$ are independent and have common distribution $P^i(\cdot)$, conditional on $P^i(\cdot)$;*
4. *there exists a random distribution $F^i(\cdot) \sim BS(c^i, F_0^i)$ such that the $T_{i,n}$ are independent and have common distribution $F^i(\cdot)$, conditional on $F^i(\cdot)$;*

5. all the $P^i(\cdot)$ and $F^i(\cdot)$ are independent.

Proof of Lemma 5.6.1. To show points (1)-(4) it suffices to note that: i) for all $i \in E$, the sequence $(L_{i,n})_{n\geq 1}$ is generated by $Dir(m^i)$-urn U_i; ii) for all $i \in E$, $(T_{i,n})_{n\geq 1}$ is generated by the $BS(c^i, F_0^i)$-system V_i; iii) the outcomes of the urns U_i, $V_{1,i}$, $V_{i,2}$, ..., for all $i \in E$ are independent of each other. To prove (5), since $(L_{i,n})_{n\geq 1}$ and $(T_{i,n})_{n\geq 1}$ are exchangeable, by the de Finetti representation theorem $P^i(\cdot) = \mathbb{P}(L_{i,1} \in \cdot|\mathcal{L}_i)$ and $F^i(\cdot) = \mathbb{P}(T_{i,1} \in \cdot|\mathcal{T}_i)$ with probability 1, where $\mathcal{L}_i$ and $\mathcal{T}_i$ are, respectively, the tail σ-fields of $(L_{i,n})_{n\geq 1}$ and $(T_{i,n})_{n\geq 1}$ (Kallenberg, 2006, Chapter 1). The thesis now follows because all σ-fields $\mathcal{L}_i$ and $\mathcal{T}_i$, $i \in E$, are independent by (1) and (2). □

Proof of Theorem 5.6.1. Take $P^i(\cdot)$ and $F^i(\cdot)$ for $i \in E$ as given by Lemma 5.6.1. Define $(\mathbf{P}, \mathbf{F})$ by letting $\mathbf{P} = (P^i(\{j\}))_{i,j\in E}$ (note that $P^i(\{i\}) = 0$ almost surely since $m^i(\{i\}) = 0$) and $\mathbf{F} = \{F^i(\cdot) : i \in E\}$. To prove the thesis it suffices to show that, conditional on $(\mathbf{P}, \mathbf{F})$, $\{(L_k, T_k)\}_{k\geq 0}$ is a Markov renewal process with characteristic couple $(\mathbf{P}, \mathbf{F})$. To do so, note that $\mathbb{P}(L_0 = l_0|(\mathbf{P}, \mathbf{F})) = 1$ by definition. Moreover, on the event $\{L_n = i, v_{i,k} = n\}$, $k \leq n$, it is

$$\begin{aligned}\mathbb{P}(L_{n+1} = j, T_n \leq t|L_{0:n}, T_{0:n-1}, (\mathbf{P}, \mathbf{F})) &= \mathbb{P}(L_{i,k} = j, T_{i,k} \leq t|(\mathbf{P}, \mathbf{F})) \\ &= P^i(\{j\})F^i(t).\end{aligned}$$

This concludes the proof. □

5.7 Generalizations of the semi-Markov beta-Stacy process

As anticipated in Remark 5.2.1, here we illustrate how the semi-Markov beta-Stacy process can be generalized to the setting where the distribution of the holding time T_k is assumed to depend on both L_k and L_{k+1}:

$$\mathbb{P}(L_{k+1} = j, T_k \leq t|L_k = i, L_{0:k-1}, T_{0:k-1}, (\mathbf{P}, \mathbf{F})) = F^{i,j}(t)P^{i,j} \tag{5.3}$$

for all $i, j \in E$ and $k \geq 0$, where now $\mathbf{F} = (F^{i,j}(\cdot) : i, j \in E, i \neq j)$, while $\mathbf{P} = (P^{i,j} : i, j \in E)$ as in Definition 5.2.1. This corresponds to the assumption that the process $\{(L_k, T_k)\}_{k\geq 0}$ evolves by first deciding which state $L_{k+1} \sim P^{L_k}(\cdot)$ will be visited after leaving the current state L_k, and only then decide how much time

$T_k \sim F^{L_k,L_{k+1}}(\cdot)$ to spend in the current state L_k. Compared to the formulation of Definition 5.2.1, the present one may be more appropriate in some applications (Barbu and Limnios, 2009).

In this new setting, the definition of the semi-Markov beta-Stacy process can be extended in two ways based on different prior assumptions. These generalizations and the process of Definition 5.3.3 are all characterized by similar reinforced urn models. These uniquely determine the predictive distributions associated to each process.

5.7.1 A first non-conjugate generalization

The most natural approach consists in defining $\mathbf{c} = (c^{i,j} : i, j \in E,\ i \neq j)$, $\mathbf{F_0} = (F_0^{i,j} : i, j \in E,\ i \neq j)$ and then substituting the symbols c^i and F^i with $c^{i,j}$ and $F^{i,j}$ in points 3 and 5 of Definition 5.3.3 (all other points remaining unchanged).

Despite its simplicity, this approach leads to a generalization of the semi-Markov beta-Stacy process which does not retain all the properties shown in the previous section. In particular, the natural generalization of Theorem 5.4.1 does not hold, as now the process in not necessarily conjugate.

This lack of conjugacy is evident from the structure of the likelihood function of $(\mathbf{P}, \mathbf{F})$ for data $S_{0:t} = i_{0:t}$, whose general form when $N(t) = n$ is

$$\begin{aligned}
\mathbb{P}\left(S_{0:t} = i_{0:t} | \mathbf{P}, \mathbf{F}\right) &= \mathbb{P}\left(L_{0:n} = l_{0:n}, T_{0:n-1} = t_{0:n-1}, T_n > l(t) | \mathbf{P}, \mathbf{F}\right) \\
&= \left[\prod_{\substack{i,j \in E \\ i \neq j}} \prod_{s=1}^{t} F^{i,j}(\{s\})^{N^{i,j,t}(\{s\})}\right] \\
&\cdot \left[\sum_{\substack{l_{n+1} \in E \\ l_{n+1} \neq l_n}} P^{l_n}(\{l_{n+1}\}) F^{l_n,l_{n+1}}((l(t), +\infty))\right] \\
&\cdot \left[\prod_{\substack{i,j \in E \\ i \neq j}} P^i(\{j\})^{M^{i,j}(t)}\right],
\end{aligned}$$

where $N^{i,j,t}$ is the number of non-terminal i-blocks of length $\leq l$ which are immediately followed by a j-block in $S_{0:t}$ ($N^{i,t}(l) = \sum_{j \in E} N^{i,j,t}(l)$).

Here, if $l(t) = 0$, i.e. $i_t \neq i_{t-1}$, the second term in the square brackets is equal to 1 (since $F^{i,j}((0, +\infty)) = 1$ for all i and j). In this case, the posterior distribution of (P, F) given the observation of the event $\{S_{0:t} = i_{0:t}\} = \{L_{0:n} = l_{0:n}, T_{0:n-1} = t_{0:n-1}\}$ is again a semi-Markov beta-Stacy process, call it $SMBS(l_{0:n}, t_{0:n-1})$, whose parameters can be obtained from obvious analogues of points 1-3 of Theorem 5.4.1.

On the other hand, if $l(t) > 0$, i.e. $i_t = i_{t-1}$, it can be shown that the posterior distribution of (P, F) given the observation of the event $\{S_{0:t} = i_{0:t}\} = \{L_{0:n} = l_{0:n}, T_{0:n-1} = t_{0:n-1}, T_n > l(t)\}$ is the mixture of semi-Markov beta-Stacy processes $SMBS((l_{0:n}, L_{n+1}), (t_{0:n-1}, T_n))$, where the mixing measure is the distribution of (L_{n+1}, T_n) given $L_{0:n} = l_{0:n}$, $T_{0:n-1} = t_{0:n-1}$, and $T_n > l(t)$.

Although the posterior distribution associated to the generalized semi-Markov beta-Stacy process is not immediately available, it is still possible to characterize its associated predictive distributions using a new reinforced urn process.

Specifically, associate every state $i \in E$ with a $Dir(m^i)$-urn U_i and every pair $(i, j) \in E \times E$, $i \neq j$, with the $BS(c^{i,j}, F_0^{i,j})$-system $V_{i,j}$ of urns $V_{i,j,1}$, $V_{i,j,2}$, $V_{i,j,3}$, and so on. Generate a sequence $\{(L_k, T_k)\}_{k\geq 0}$ as follows. First, set $L_0 = l_0$. Then, for all $k \geq 0$, generate L_{k+1} from U_{L_k} and, independently, T_k from $V_{L_k, L_{k+1}}$. Lastly, denote with $S = (S_t)_t \geq 0$ the process with state space E induced by $\{(L_k, T_k)\}_{k\geq 0}$ as in Section 5.6.

Additionally, for all $(i, j) \in E \times E$, $i \neq j$, let $v_{i,j,0} = -1$ and, for all integers $n \geq 1$, let $v_{i,j,n} = \inf\{k > v_{i,j,n-1} : L_k = i\}$ be the time the sequence $(L_k)_{k\geq 0}$ performs its n-th transition from the state i to the state j. Note that $v_{i,n}$, the time of the n-th visit to i, is related to the $v_{i,j,n}$, $j \neq i$, by $v_{i,n} = \min\{v_{i,j,k} : j \neq i,\ k \leq n,\ v_{i,j,k} > v_{i,n-1}\}$. We will consider the following strengthening of the recurrence condition of Equation 5.2:

$$\mathbb{P}\left(\bigcap_{\substack{(i,j)\in E\times E\\ i\neq j}} \bigcap_{n=1}^{+\infty} \{v_{i,j,n} < +\infty\}\right) = 1. \tag{5.4}$$

This not only implies that $(S_t)_{t\geq 0}$ is recurrent, but also that it performs every allowable transition an infinite number of times with probability 1. Hence, the sequences $(L_{i,k})_{k\geq 0} = (L_{v_{i,k}+1})_{k\geq 0}$ and $(T_{i,j,k})_{k\geq 0} = (T_{v_{i,j,k}})_{k\geq 0}$ are infinite with probability 1.

Importantly, under the condition of Equation (5.4), Theorem 5.6.1 still holds. In fact, proceeding as in the proof of Lemma 5.6.1, under condition (5.4) it can be shown that: i) the arrays of random variables $\{(L_{i,n}, T_{i,j,n} : j \neq i)\}_{n\geq 0}$, $i \in E$,

are independent of each other; ii) the sequences $(L_{i,n})_{n\geq 0}$, $(T_{i,j,n})_{n\geq 0}$, $i \neq j$, are all independent of each other; iii) for all $i \in E$, the $L_{i,n}$ are independent and identically distributed as $P^i(\cdot)$ for some random $P^i(\cdot) \sim Dir(m^i)$; iv) for all $i \neq j$, the $T_{i,j,n}$ are independent and identically distributed as $F^{i,j}(\cdot)$ for some random $F^{i,j}(\cdot) \sim BS(c^{i,j}, F_0^{i,j})$; and v) all the $P^i(\cdot)$ and $F^{i,j}(\cdot)$, $i \neq j$, are independent of each other.

Consequently, letting $\mathbf{P} = (P^i(\{j\}))_{i,j\in E}$ and $\mathbf{F} = \{F^{i,j}(\cdot) : i, j \in E, i \neq j\}$, it is $\mathbb{P}(L_0 = l_0|(\mathbf{P}, \mathbf{F})) = 1$. Moreover, on the event $\{L_n = i, v_{i,k} = v_{i,j,h} = n\}$ with $h \leq k \leq n$, it is

$$\begin{aligned}\mathbb{P}(L_{n+1} = j, T_n \leq t|L_{0:n}, T_{0:n-1}, (\mathbf{P}, \mathbf{F})) &= \mathbb{P}(L_{i,k} = j, T_{i,j,h} \leq t|(\mathbf{P}, \mathbf{F})) \\ &= P^i(\{j\})F^{i,j}(t),\end{aligned}$$

as desired.

5.7.2 An alternative conjugate generalization

To arrive at an alternative generalization, we consider the following approach. First, note that Equation 5.3 can be equivalently expressed as

$$\mathbb{P}(L_{k+1} = j, T_k = t|L_k = i, L_{0:k-1}, T_{0:k-1}, (\mathbf{P}, \mathbf{F})) = F^i(\{t\})P_t^{i,j} \tag{5.5}$$

where $F^i(t) = \sum_{j\neq i} F^{i,j}(t)P^{i,j}$ and $P_t^{i,j} = F^{i,j}(t)P^{i,j}/F^i(t)$, where now $\mathbf{P} = (P_t^{i,j} = P_t^i(\{j\}) : i, j \in E, t \geq 1)$, while $\mathbf{F} = (F^i(\cdot) : i \in E)$ as in Definition 5.2.1. In this formulation, the process $\{(L_k, T_k)\}_{k\geq 0}$ evolves by first deciding the time $T_k \sim F^{L_k}(\cdot)$ to spend in the current state L_k subsequently deciding the next state $L_{k+1} \sim P_{T_k}^{L_k}(\cdot)$.

From this perspective, Definition 5.3.3 can be generalized by letting $\mathbf{m} = (m_t^i : i \in E, t \geq 1)$ be a family of measures on E and then supposing that the $P_t^i(\cdot)$ are independent $Dir(m_t^i)$ processes on E for all $i \in E$ and $t \geq 1$ (all other assumptions remaining as is).

Contrary to the previous case, this generalization of the semi-Markov beta-Stacy process is easily seen to be conjugate. In fact, an immediate generalization of Theorem 5.4.1 can be obtained by noting that the likelihood function of $(\mathbf{P}, \mathbf{F})$ for data

$S_{0:t} = i_{0:t}$ such that $N(t) = n$ now takes the form

$$\mathbb{P}\left(S_{0:t} = i_{0:t} | \mathbf{P}, \mathbf{F}\right) = \left[\prod_{i \in E} \prod_{s=1}^{t} F^i(\{s\})^{N^{i,t}(\{s\})} F^i((l(t), +\infty))^{I\{i=i_n\}}\right] \cdot \left[\prod_{\substack{i,j \in E \\ i \neq j}} \prod_{s=1}^{t} P_s^i(\{j\})^{N^{i,j,t}(\{s\})}\right]$$

(note that $M^{i,j}(t) = \sum_{s=1}^{t} N^{i,j,t}(\{s\})$ for all $i \neq j$). Thus, the posterior distribution of $(\mathbf{P}, \mathbf{F})$ given $S_{0:t} = i_{0:t}$ is a $SMBS(\mathbf{m}_*, \mathbf{c}_*, \mathbf{F}_*)$, where $\mathbf{c}_*$ and $\mathbf{F}_*$ are defined as in Theorem 5.4.1, while $\mathbf{m}_* = (m^i_{*,s} : i \in E, s \geq 1)$ is obtained by letting $m^i_{*,s}(\cdot) = m^i_s(\cdot) + N^{i,j,t}(\{s\})$ for all $i \in E$ and $s \geq 1$.

As before, this generalization of the semi-Markov beta-Stacy process can also be characterized by another reinforced urn process. Specifically, associate every $i \in E$ with a $BS(c^i, F^i_0)$-system V_i as in Section 5.6 and every couple $(i, t) \in E \times \{1, 2, 3 \ldots\}$ with a $Dir(m^i_t)$-urn $U_{i,t}$. Suppose that $\{(L_k, T_k)\}_{k \geq 0}$ is generated first by letting $L_0 = l_0$ and then by iteratively generating T_k from V_{L_k} and L_{k+1} from U_{L_k, T_k} for all $k \geq 0$. In this set up, a generalization of Theorem 5.6.1 can be shown to hold under an appropriate strengthening of the recurrence condition of Equation 5.2. In particular, it suffices to require that every urn is visited infinitely often with probability one.

5.8 Simulation study

To illustrate the semi-Markov beta-Stacy process in action, we conducted a simulation study based on the textile factory scenario of Example 5.2.2.

5.8.1 Description of the simulation study

Following Barbu and Limnios (2009, Sections 4.3), we generated a single realization $s_{0:1,000}$ from the semi-Markov process $(S_t)_{t \geq 0}$ describing the day-by-day status of the factory from day 0 to day 1,000. The law of this process was determined by assuming that: i) $S_0 = 1$ (so the factory begins fully functional); ii) the transition

matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0.95 & 0 & 0.05 \\ 1 & 0 & 0 \end{bmatrix}; \tag{5.6}$$

iii) $F^1(\cdot)$ is the geometric distribution $F^1(\{t\}) = p(1-p)^{t-1}$, $t \geq 1$, with parameter $p = 0.8$; iv) $F^2(\cdot)$ is the first-type discrete Weibull distribution $F^2(t) = 1 - q^{t^k}$, $t \geq 1$, of Nakagawa and Osaki (1975) with parameters $q = 0.3$ and $k = 0.5$ (when $k = 1$, this distribution reduces to the geometric distribution with parameter $1 - q$); v) $F^3(\cdot)$ is the first-type discrete Weibull distribution with parameters $q = 0.6$ and $k = 0.9$. The observed sequence $s_{0:1,000}$ was considered as data to perform posterior inferences.

5.8.2 Prior specification

We assign a semi-Markov beta-Stacy prior distribution $SMBS(\mathbf{m}, \mathbf{c}, \mathbf{F_0})$ to the data-generating characteristic couple $(\mathbf{P}, \mathbf{F})$. We consider the measures $m^1(\cdot)$, $m^2(\cdot)$, and $m^3(\cdot)$ on $E = \{1, 2, 3\}$ determined by the conditions $m^i(\{1, 2, 3\}) = m^1(\{2\}) = m^2(\{1\}) = m^2(\{3\}) = m^3(\{1\}) = 1$ for all $i \in E$ (in particular, this implies that both $P^{2,1}$ and $P^{2,3}$ are marginally uniformly distributed over $(0, 1)$). For all $i = 1, 2, 3$, $F_0^i(\cdot)$ will be the geometric distribution with parameter $p = 0.3$ (a prior assumption clearly incompatible with the data-generating mechanism). For all $i \in E$, we consider $c^i(t) = c$ for all $t \geq 1$ and some constant $c > 0$, successively considering the values $c = 0.1$, 1, and 10.

5.8.3 Posterior distributions

Figure 5.2 shows the plots of the posterior mean of $F^2(\cdot)$, together with a sample of 500 samples from the corresponding distribution. Posterior distributions were obtained from Theorem 5.4.1 using data $s_{0:M}$ with $M = 0$ (so the posterior coincides with the prior), $M = 100$, or $M = 1000$ (so whole simulated path is used). For comparison, the figure also reports the data-generating distribution of the holding-times of the state 2, i.e. of the time elapsed until either the tank is repaired or the factory has to stop after a failure.

Figure 5.2 highlights how the posterior distribution obtained from the semi-Markov beta-Stacy prior is able to recover the underlying data-generating distribu-

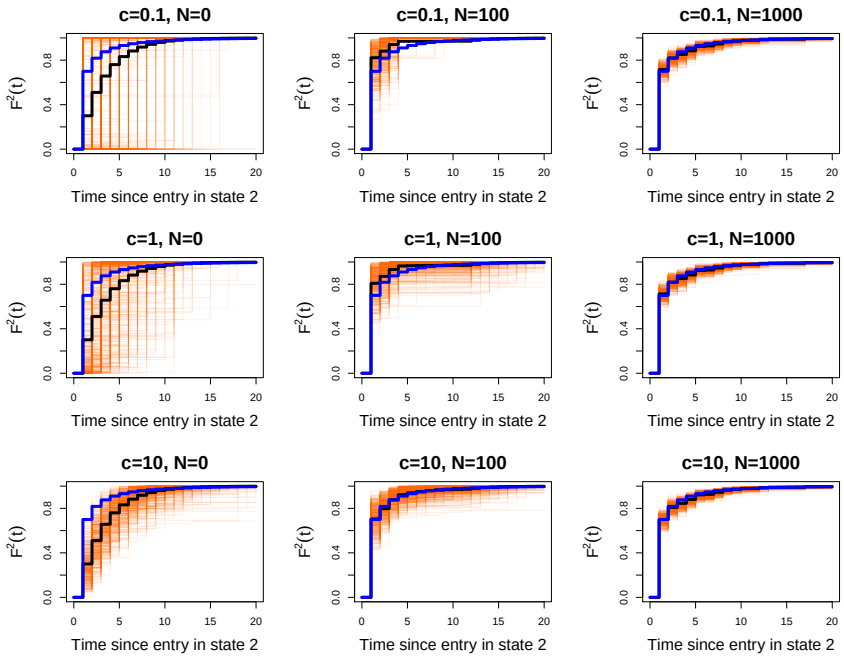


Figure 5.2: Plot of the posterior distribution of $F^2(\cdot)$ for the semi-Markov process priors of Section 5.8. Results are shown for different values of: i) the prior concentration parameter c, which specifies the weight assigned to the prior centering distributions $F_0^2(\cdot)$; ii) the length N of the observation period during which data $S_{0:N}$ is collected (if $N = 0$, the posterior distribution coincides with the prior). Blue lines: values of the true data-generating distribution $F^2(\cdot)$ (see Section 5.8.1). Black lines: posterior mean of $F^2(\cdot)$. Orange lines: graph of 500 samples from the posterior distribution of $F^2(\cdot)$.

tion by flexibly adapting to the observations, even when these deviate from prior assumptions. This is true both for data reflecting a short ($M = 100$) or long ($M = 1,000$) period of observation. The figure also highlights the impact of the concentration parameters c. As this increases, the dispersion of the distribution of $F^2(\cdot)$ around its mean decreases.

5.8.4 Predictive distributions

Figure 5.3 reports the estimates of the predictive distributions $P_h(j) = \mathbb{P}(S_{1,000+h} = j | S_{0:1,000} = s_{0:1,1000})$ obtained from the semi-Markov beta-Stacy prior with $c = 1$ for

all $h = 1, \ldots, 100$ and all $j = 1, 2, 3$. These were obtained by simulating 10^5 future paths $(S_{1,000+h})_{h=1,\ldots,100}$ conditional on the past observation of $S_{0:1,000} = s_{0:1,000}$ by sampling from the reinforced semi-Markov kernels of Corollary 5.5.1. Then, $P_h(j)$ was estimated as the proportion of simulations in which $S_{1,000+h} = j$.

Figure 5.3 shows how the the $P_h(j)$ adapt over time as h increases for all $j = 1, 2, 3$, whose values stabilize in the long run. Specifically, for large h the vector $(P_h(1), P_h(2), P_h(3))$ remain close to the limiting distribution (ν_1, ν_2, ν_3) of the data-generating semi-Markov process. This is obtained from Proposition 3.9 of Barbu and Limnios (2009) as $\nu_j = e_j m_j / \sum_{i=1}^{3} e_i m_i$, where $(e_1, e_2, e_3) = (\frac{1}{2.05}, \frac{1}{2.05}, \frac{0.05}{2.05})$ is the equilibrium distribution of the transition matrix $\mathbf{P}$ in Equation 5.6, while $m_j = \sum_{t=0}^{+\infty}(1 - F^j(t))$ is the expected sojourn time in the state j.

5.9 Concluding remarks

In this paper we introduced the semi-Markov beta-Stacy process, a Bayesian nonparametric process prior for semi-Markov models, and some related generalizations. Each was characterized from a predictive perspective by "piecing together" different reinforced urn models characterizing simpler processes.

This approach is conceptually valuable, as it provides a fresh strategy for the specification of Bayesian nonparametric models for the prediction of complex processes. Importantly, as previously noted by Muliere et al. (2003), reinforced stochastic processes can be used to perform predictions from a Bayesian nonparametric perspective without requiring knowledge of difficultly obtained aspects of the prior or posterior distributions (Ghosal and van der Vaart, 2017).

The semi-Markov beta-Stacy may be amenable to more generalization then the ones considered here by modifying its underlying reinforced urn process. First, each extracted ball may be reinforced by a fixed or random amount of multiple balls of the same or different colors, akin as in Muliere et al. (2006). This could allow a finer control of the level of uncertainty attached to the urns' initial composition, i.e. to the centering distribution of the prior (Arfè et al., 2018).

Second, a form of dependence across different components of the prior may be introduced by reinforcing urns other than the one from which a ball was extracted. This form of interaction among urns could lead to interesting models in which observations provide indirect information about distributions that have not generated

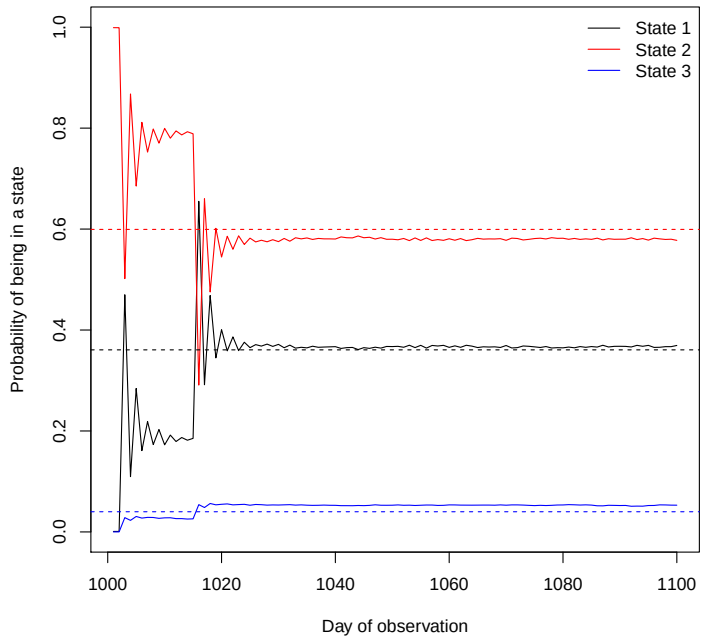


Figure 5.3: Plot of the predictive probabilities $P_h(j) = \mathbb{P}(S_{1,000+h} = j | S_{0:1,000} = s_{0:1,1000})$ obtained from the semi-Markov beta-Stacy process of Section 5.8 with $c = 1$ for all $h = 1, \ldots, 100$. The value $P_h(j)$ is the probability that the factory will be in state $j = 1, 2, 3$ after h days in the future given its past history $S_{0:1000}$. The black, red, and blue lines are, respectively, the values of $P_h(1)$, $P_h(2)$, and $P_h(3)$. The dashed lines represent the limiting distribution of the underlying data-generating semi-Markov process.

them directly (Paganoni and Secchi, 2004; Muliere et al., 2005).

From a more applied perspective, we are investigating different ways to exploit the semi-Markov beta-Stacy process in more complex Bayesian non-parametric models based on semi-Markov processes. In particular, we are implementing a regression model in which the distribution of the holding times and the transition matrices depend on a vector of covariates. As in Arfè et al. (2018), this is done by letting the initial composition of the urns be a function of both the covariates and some additional parameters, which are then assigned their own prior distribution. Such model could be used for the analysis of multi-stage diseases in medical studies (Barbu et al., 2004; Mitchell et al., 2011).

Additionally, we are applying the semi-Markov beta-Stacy process to perform inference and predictions in Hidden Semi-Markov Models (HSMMs), in which the sequence of visited states is observed only indirectly (Barbu and Limnios, 2009, Chapter 6). As a specific application, we are developing a novel approach for changepoint analysis in which the state of a semi-Markov process represents the latent regimen of a time series (Smith, 1975; Muliere and Scarsini, 1985; Ko et al., 2015; Peluso et al., 2018).

Chapter 6

Bayesian optimality of testing procedures for survival data in the non-proportional hazards setting

With Brian Alexander and Lorenzo Trippa.
Submitted manuscript with invited revisions.
ArXiv manuscript: `https://arxiv.org/abs/1902.00161`

6.1 Introduction

Researchers often use data generated by exploratory clinical studies to specify the protocol of randomized confirmatory phase III trials. Data predictive of the confirmatory trial outcomes, including early estimates of treatment effects, are used to choose the primary endpoints (Gómez et al., 2014), the sample size (Lindley, 1997), the target populations (Lee and Wason, 2018), and other aspects of the study design (Brody, 2016). Still, in most cases prior information is not used to specify in the protocol, as mandated by regulatory agencies, which hypothesis testing procedure will be used in the final analyses to provide evidence of treatment effects. Agencies such as the U.S. Food and Drug Administration require the control of Type I and II errors at acceptable, pre-specified rates (US Food and Drug Administration, 1998).

In Phase III trials, standard tests, such as Mantel's log-rank, are often selected even for studies where prior data suggests their underlying assumptions will be

violated (Royston and Parmar, 2013; Alexander et al., 2018). For survival endpoints, methods related to the log-rank test are prevalent. Asymptotically, this is the most powerful test with a proportional hazards alternative (Fleming and Harrington, 2011). However, the proportional hazards assumption is often violated in practice, contributing to false-negative findings (Royston and Parmar, 2013), invalidating sample size calculations (Barthel et al., 2006), and affecting interim analyses (van Houwelingen et al., 2005).

Data from early-stage studies can inform about deviations from the assumption of proportional hazards, suggesting the use of alternative methods (Royston and Parmar, 2013). Several extensions and alternatives are available to replace Mantel's test, such as weighted (Fleming and Harrington, 2011) or adaptive log-rank tests (Yang and Prentice, 2010), and restricted mean survival tests (Royston and Parmar, 2013). Several of these procedures identify the most powerful test against specific alternatives, which may not be representative of available estimates from early stage analyses. Moreover, their optimality typically holds in a large-sample sense (e.g. in the local limit for weighted log-rank tests; Fleming and Harrington 2011).

We develop a statistical test to detect treatments effects in late-stage trials, accounting for deviations from the proportional hazards assumption indicated by early-phase studies (e.g. phase II trials). The proposed test does not belong to the weighted log-rank family or other common classes of tests. Starting from decision theory principles (Robert, 2007), we derive it as the solution to the following constrained decision problem (Ventz and Trippa, 2015): conditional on early-stage data, the test maximizes the predicted finite-sample power among all tests which control the frequentist Type I error rate of the late-stage study at a fixed α level. More precisely, the test maximizes the Bayesian predictive probability that the null hypothesis will be correctly rejected at the end of the confirmatory trial. The test therefore provides a useful benchmark for other procedures applicable in presence of non-proportional hazards.

As a motivating example we consider the analysis of a randomized trial with delayed treatment effects on survival outcomes. This is a characteristic which occurs when the treatment requires an induction period before it starts to exert therapeutic effects. When treatment effects are delayed, the hazard functions are not proportional and they separate across arms only later during follow-up (Fine, 2007). Initially overlapping survival curves (c.f. Figure 6.1a) are well documented in trials

of cancer immunotherapies (Chen, 2013; Alexander et al., 2018). They can also be observed in other settings, such as in studies of breast cancer (Mehta et al., 2012) and melanoma (Robert et al., 2015) chemotherapies.

6.2 Example

We consider data on the survival times of the 361 patients with head and neck carcinomas that participated in CheckMate 141 study (Ferris et al., 2016), a Phase III trial that randomized patients to receive nivolumab, a novel cancer immunotherapy, or standard of care (SOC) in a 2:1 ratio. We reconstructed the individual-level data of this trial from Figure 1a of Ferris et al. (2016) by means of the DigitizeIt (TM) software (version 2.2) and the data extraction method of Guyot et al. (2012). Figure 6.1a shows the resulting Kaplan-Meier curves, which compare survival probabilities between the two study arms. These do not clearly separate in the initial 3-4 months of follow-up, a signal of delays in the treatment effects.

6.3 Planning a late-stage trial

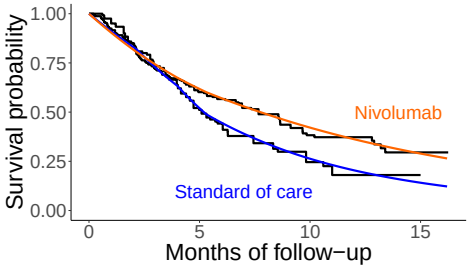
We plan a late-stage randomized trial with a survival end-point and a sample size of n patients. This will generate data $x = (t, d, a)$ to test if the treatment has positive effects on the primary outcome. Here, $t = (t_1, \ldots, t_n)$ are the observed follow-up times, $d = (d_1, \ldots, d_n)$ are the corresponding censoring indicators ($d_i = 1$ if t_i is censored, while $d_i = 0$ if an event was observed), and $a = (a_1, \ldots, a_n)$ are the study arm indicators ($a_i = 0$ or $a_i = 1$ if the i-th patient is randomized to the control or treatment arm). Patients are assigned to arms with a fixed randomization probability. We assume that censoring times are *non informative* in the sense of Heitjan and Rubin (1991) and independent of treatment assignment.

For design purposes, we specify a model for the distribution that will generate the data x. This is described by a density $p_\theta(x)$ that depends on a parameters vector $\theta \in \Theta$. Here θ may be infinite-dimensional if the model is semi- or non-parametric. Typically, $p_\theta(x)$ will have the form

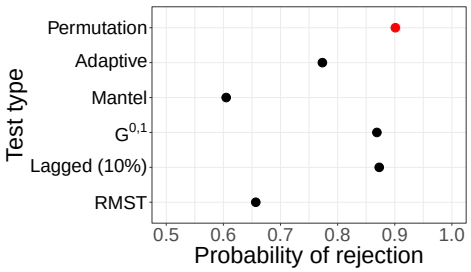
$$p_\theta(x) = \prod_{i=1}^{n} r^{a_i}(1-r)^{1-a_i} h_{a_i}(t_i;\theta)^{1-d_i} S_{a_i}(t_i;\theta) g_i(t_i)^{d_i} G_i(t_i)^{1-d_i}, \qquad (6.1)$$

Figure 6.1: Panel a, reconstructed Kaplan-Meier curves from the CheckMate 141 trial and posterior estimates obtained from the piecewise exponential model (Section 6.6). Panel b, Monte Carlo estimates of the rejection probability of selected tests (Section 6.7.1). Panel c, results of the robustness analysis (Section 6.7.2). Legend: permutation, maximum-BEP test of Section 6.5 based on the piecewise exponential model (highlighted in red); adaptive, adaptive log-rank test of Yang and Prentice (2010); mantel, classical Mantel's log-rank test; $G^{0,1}$, Fleming-Harrington weighted log-rank test; lagged, lagged-log rank that ignores the first 10% of observed follow-up times (Zucker and Lakatos, 1990), RMST, test of the difference in restricted mean survival times (Huang and Kuan, 2018).

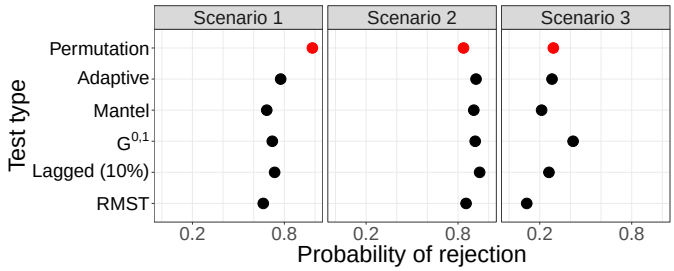
(a)



(b)



(c)



where: $r \in (0,1)$ is the probability of assignment to arm $a = 1$, $h_a(t;\theta) > 0$ is the hazard function of arm $a = 0, 1$ (for example, in the exponential model, $h(t;\theta) = \theta_a$, $\theta = (\theta_0, \theta_1) \in \Theta = (0, +\infty)^2$); $S_a(t;\theta) = \exp\left(-\int_0^t h_a(s;\theta)ds\right)$ is the corresponding survival function; finally, $g_i(t)$ and $G_i(t)$ are the density and (left-continuous) censoring function of the i-th patient. Here, the censoring mechanism is taken as known, a common assumption when planning new experiments (Chow et al., 2007). We will later discuss that this assumption is not used in the development of the proposed testing procedure.

We consider the non-parametric null hypothesis $H_0 : P \in \mathcal{P}_0$, where P is the true data-generating distribution of x (i.e. $P(A)$ is the probability that $x \in A$) and $\mathcal{P}_0$ is the class of all distributions which are invariant with respect to permutations of the treatment arm assignments. Hence, $P \in \mathcal{P}_0$ if its likelihood function $p(x)$ is such that $p(t, d, a) = p(t, d, a')$ for all a' obtained by permuting the elements of a.

The alternative hypothesis is instead defined using model (6.1) as H_1: P has density $p_\theta(x)$ for some $\theta \in \Theta_1$, where Θ_1 is a subset of Θ. For example, Θ_1 may include all θ such that $h_0(t;\theta) \neq h_1(t;\theta)$, or such that the median of $S_1(t;\theta)$ is greater than that of $S_0(t;\theta)$, or such that the restricted mean survival in arm $a = 1$ is greater than in arm $a = 0$ (Royston and Parmar, 2013).

According to this definition of the null hypothesis, regardless of whether the model $p_\theta(x)$ is correct or not, when treatment has no effect the treatment assignments $a_1, \ldots, a_n$ provide no information about the follow-up times t and censoring indicators d. Hence, the distribution of the data does not change if these are arbitrarily permuted (c.f. Fisher, 1935; Dawid, 1988; Good, 2006; Pesarin and Salmaso, 2010).

This definition covers distributions in which the observations $(t_1, d_1, a_1), \ldots, (t_n, d_n, a_n)$ from individual patients cannot be considered independent and identically distributed. For example, this may happen when recruiters selectively enroll patients in the trial based on interim analyses or results from other studies published during the enrollment period, or when treatment effects are confounded by trends in latent covariates, amendments of inclusion-exclusion criteria, and improvements in adjuvant therapies (Tamm and Hilgers, 2014). In such cases, if treatment has no effects we may still expect $p(x)$ to remain invariant if the treatment arm indicators are permuted.

It is now necessary to choose which α-level test $\varphi(x)$ should be used in the late-

stage trial. A *(randomized) test* of H_0 is a function $\varphi(x) \in [0,1]$ such that if data x is observed, then H_0 is rejected with probability $\varphi(x)$ (Lehmann and Romano, 2006). A test is *non-randomized* if it can only attain the values 0 and 1 (only non-randomized tests are used in practice, but here we also consider randomized tests because of their analytic advantages). The expected value $E_P[\varphi(x)] = \int_{\mathcal{X}_n} \varphi(x) dP(x)$ is equal to the probability of rejecting H_0 with data generated from the distribution P. If $\alpha \in (0,1)$ and $E_P[\varphi(x)] \leq \alpha$ for all $P \in \mathcal{P}_0$, then $\varphi(x)$ is said to have *level* α.

6.4 Bayesian expected power

Different α-level tests are usually compared with respect to their *power functions* $\pi_\varphi(\theta) = \int \varphi(x) p_\theta(x) dx$ or its asymptotic approximations. If $\varphi_1(x)$ and $\varphi_2(x)$ are two α-level tests for H_0 versus the simple alternative $H_1 : \theta = \theta_1$, for some fixed $\theta_1 \in \Theta_1$, then $\varphi_1(x)$ is preferred to $\varphi_2(x)$ if $\pi_{\varphi_1}(\theta_1) \geq \pi_{\varphi_2}(\theta_1)$. Such comparisons are difficult for composite alternative hypotheses. In fact, uniformly most powerful α-level tests, i.e. tests achieving the maximum power across all alternative models $\theta_1 \in \Theta_1$, do not necessarily exist (Lehmann and Romano, 2006).

To address this problem, some authors proposed to compare tests with respect to their *average power*. Specifically, the average power of a test $\varphi(x)$ is $\int_{\Theta_1} \pi_\varphi(\theta) p(\theta) d\theta$, where $p(\theta)$ is a distribution weighting each value of $\theta \in \Theta$ based on pre-experimental information (Spiegelhalter and Freedman, 1986; O'Hagan et al., 2005). With this metric, two tests are always comparable. Additionally, α-level tests maximizing the average power always exist, although these may be randomized (Chen et al., 2007).

To allow data $x_e = (t_e, d_e, a_e)$ from an early-stage trial to inform comparisons between tests, we consider a data-dependent prior $p(\theta|x_e)$. Several approaches have been proposed to incorporate historical data in a prior distribution, including *power priors* (Ibrahim et al., 2015), *meta-analytic priors* (Schmid et al., 2016), and *commensurate priors* (Hobbs et al., 2011). For simplicity, we define $p(\theta|x_e)$ as the posterior distribution $p(\theta|x_e) \propto L(\theta; x_e) p(\theta)$, where, letting n_e be the early-stage trial sample size,

$$L(\theta; x_e) = \prod_{i=1}^{n_e} h_{a_{e,i}}(t_{e,i}; \theta)^{1-d_{e,i}} S_{a_{e,i}}(t_{e,i}; \theta), \tag{6.2}$$

while $p(\theta)$ is a prior distribution on Θ whose choice depend on the specific application

context. In doing so, we implicitly assume identical treatment effects and survival distributions in the early- and late-stage trials.

Extending the average power approach, the *Bayesian expected power (BEP)* of $\varphi(x)$ is defined as

$$BEP_\varphi = \int_{\Theta_1} \pi_\varphi(\theta) p(\theta|x_e) d\theta, \tag{6.3}$$

a concept first introduced by Brown et al. (1987) and "rediscovered" by several authors (Liu, 2018). It is simple to observe that $BEP_\varphi = \Pr(\varphi(x) \text{ rejects } H_0 \text{ and } \theta \in \Theta_1|x_e)$, the probability, conditional on the early-stage data, that $\varphi(x)$ will correctly reject H_0 at end of the late-stage trial. This is often called the *probability of success* of the trial (Liu, 2018).

From the point of view of decision theory (Robert, 2007), the BEP is the expected value of the utility function $u(\theta, \varphi, x) = I\{\theta \in \Theta_1\}\varphi(x)$ (if H_1 holds, then the utility increases with the probability $\varphi(x)$ of rejecting H_0). Indeed,

$$BEP_\varphi = \int\int u(\theta, \varphi, x) p_\theta(x) p(\theta|x_e) dx d\theta. \tag{6.4}$$

The problem of choosing which test to apply in the late-stage trial can thus be stated as a constrained maximization problem (Ventz and Trippa, 2015): among α-level tests we optimize the BEP.

6.5 Tests maximizing the expected power

We identify an α-level test with maximum Bayesian expected power. Explicit expressions have been obtained for the case where the set $\mathcal{P}_0$ which defines the null hypothesis ($H_0 : P \in \mathcal{P}_0$) is finite (Chen, 2013). Instead, our choice of H_0 includes all distributions that are invariant with respect to permutations of treatment assignment $a_1, \ldots, a_n$. We show that the maximum-BEP test is a *permutation test*. This is obtained by computing or approximating the distribution of real-valued *test statistic* $T(x)$ across all permutations of the treatment assignments, while the values of the follow-up times t and censoring indicators d are kept fixed at the observed values.

To be more formal, for each permutation σ of $(1, \ldots, n)$, we denote with $a_\sigma = (a_{\sigma(1)}, \ldots, a_{\sigma(n)})$ the vector obtained by re-ordering the elements of $a = (a_1, \ldots, a_n)$ according to σ. Moreover, if $T(x)$ is any real-valued statistics, for each $x = (t, d, a)$

we let $T^{(1)}(x) \leq \cdots \leq T^{(n!)}(x)$ be the ordered values of $T(t, d, a_\sigma)$ as σ varies across all $n!$ permutations.

The α-level permutation test $\varphi(x)$ of H_0 based on the test statistic $T(x)$ can now be defined as follows. First, let $k_\alpha = n! - \lfloor \alpha n! \rfloor$, so that, for each x, $T^{(k_\alpha)}(x)$ is the $(1-\alpha)$-level quantile of $T^{(j)}(x)$ for $j = 1, \ldots, n!$. Second, let $M^+(x) = \sum_{j=1}^{n!} I\{T^{(j)}(x) > T^{(k_\alpha)}(x)\}$ and $M^0(x) = \sum_{j=1}^{n!} I\{T^{(j)}(x) = T^{(k_\alpha)}(x)\}$ be the number of $T^{(j)}(x)$'s greater or equal to $T^{(k_\alpha)}(x)$, respectively. Then, the permutation test $\varphi(x)$ is defined by letting $\varphi(x) = 1$ when $T(x) > T^{(k_\alpha)}(x)$, $\varphi(x) = 0$ when $T(x) < T^{(k_\alpha)}(x)$, and $\varphi(x) = (\alpha n! - M^+(x))/M^0(x) < 1$ when $T(x) = T^{(k_\alpha)}(x)$. This satisfies the equality $E_P[\varphi(x)] = \alpha$ for all $P \in \mathcal{P}_0$ (Lehmann and Romano, 2006, Theorem 15.2.1).

Proposition 6.5.1. Let $\varphi(x)$ be the α-level permutation test of $H_0 : P \in \mathcal{P}_0$ based on the test statistic $T(x) = q(x)$, where $q(x)$ is the density of $Q \notin \mathcal{P}_0$, $Q(A) = \int_A q(x)d\mu(x)$ for every measurable A, and μ is invariant with respect to permutations σ assignments a_1, …, a_n. If $\varphi'(x)$ is another α-level test of H_0, then $E_Q[\varphi'(x)] \leq E_Q[\varphi(x)]$, i.e. $\varphi(x)$ has higher power under the alternative $H_1 : P = Q$.

Proof. Let $\mathcal{P}_\mu \subseteq \mathcal{P}_0$ be the set of all distributions dominated by μ that are invariant with respect to permutations of treatment assignment (a non-empty set, since it includes $q'(t, d, a) = \sum_\sigma q(t, d, a_\sigma)/n!$). By Theorem 2 of Lehmann and Stein (1949), for every test $\varphi'(x)$ such that $E_P[\varphi'(x)] \leq \alpha$ for all $P \in \mathcal{P}_\mu$, $\varphi(x)$ guarantees $E_Q[\varphi'(x)] \leq E_Q[\varphi(x)]$. Now, if $\varphi'(x)$ is an α-level test of H_0, then $E_P[\varphi'(x)] \leq \alpha$ for all $P \in \mathcal{P}_\mu$ and therefore $E_Q[\varphi'(x)] \leq E_Q[\varphi(x)]$. □

To proceed, let $P(H_1|x_e) = \int_{\Theta_1} p(\theta|x_e)d\theta > 0$ be the prior probability of the alternative hypothesis H_1. Also,

$$q(x) = \int_{\Theta_1} p_\theta(x) \frac{p(\theta|x_e)}{P(H_1|x_e)} d\theta \tag{6.5}$$

is the density of the predictive distribution of x conditional on $\theta \in \Theta_1$ based on the early-stage data x_e, and $Q(A) = \int_A q(x)d\mu(x)$. Here, we assume that all densities $p_\theta(x)$ are taken with respect to the same dominating measure μ.

Proposition 6.5.2. For any test $\varphi(x)$ of H_0 we have $BEP_\varphi = E_Q[\varphi(x)]$, the power of $\varphi(x)$ against the simple alternative $H_1 : P = Q$. Consequently, a test $\varphi(x)$ maximizes the BEP among all α-level tests if and only if it maximizes the power $E_Q[\varphi(x)]$ among all α-level tests.

Proof. By Fubini's theorem, the BEP of a test $\varphi(x)$ can be written as

$$\begin{aligned} BEP_\varphi &= \int_{\Theta_1} \pi_\varphi(\theta) p(\theta|x_e) d\theta \\ &= \int_{\Theta_1} \left[\int \varphi(x) p_\theta(x) d\mu(x) \right] p(\theta|x_e) d\theta \\ &= \int \varphi(x) \left[\int_{\Theta_1} p_\theta(x) p(\theta|x_e) d\theta \right] d\mu(x) \\ &= \int \varphi(x) q(x) d\mu(x) \cdot P(H_1|x_e) \\ &= E_Q[\varphi(x)] \cdot P(H_1|x_e). \end{aligned}$$

□

Without loss of generality, to derive a maximum-BEP test we assume that μ is invariant with respect to permutations of the treatment assignments.

Using Propositions 6.5.1 and 6.5.2, we can now prove that it is possible to construct a maximum-BEP test which depends on the data x only through the *marginal likelihood*

$$m(x) = \int_{\Theta_1} L(\theta; x) p(\theta|x_e) d\theta. \tag{6.6}$$

Since $m(x)$ does not depend on the censoring distribution functions $G_i(t)$ which appear in Equation 6.1, the censoring mechanism is irrelevant to identify the optimal test. Note, however, that the censoring mechanism still determines the BEP.

Proposition 6.5.3. Given the early-stage data x_e, the α-level permutation test based on the marginal likelihood $T(x) = m(x)$ maximizes the BEP among all α-level tests of H_0.

Proof. By Proposition 6.5.1, the α-level permutation test $\varphi'(x)$ based on the test statistic $T'(x) = q(x)$ maximizes the power $E_Q[\varphi(x)]$ among all α-level tests of H_0. By Proposition 6.5.2, $\varphi'(x)$ has maximum BEP among all α-level tests of H_0. It now suffices to show that $\varphi'(x) = \varphi(x)$ for all x such that $q(x) > 0$, where $\varphi(x)$ is the α-level permutation test based on $T(x) = m(x)$. To do so, note that, by Equation 6.1, if $q(x) > 0$, then $m(x) > 0$ as well, and the ratio $q(x)/m(x)$ is invariant with respect to permutations of the treatment arm assignments. Indeed, censoring times and treatment assignments are independent. The thesis now follows because $q(t, d, a_\sigma) \propto m(t, d, a_\sigma)$ for all permutations σ.

□

Since randomized tests are not used in applications, we will consider the non-randomized version $\varphi'(x) = I\left\{m(x) > m^{(k_\alpha)}(x)\right\}$ of the test $\varphi(x)$ from Theorem 6.5.3. Since $\varphi'(x) \leq \varphi(x)$, $\varphi'(x)$ is α-level for H_0, although it may not achieve the maximum BEP. Nevertheless, $\varphi'(x)$ still provide a useful benchmark for other tests of H_0, as its BEP is close to optimal for large n. In fact, in the Appendix, Proposition 6.10.1, we show that, under mild conditions, $0 \leq BEP_\varphi - BEP_{\varphi'} \leq f(\alpha, r, n)$, where the bound is a known function such that $f(\alpha, r, n) \to 0$ as $n \to +\infty$ for all fixed levels α and randomization probabilities r. In such cases, a moderate size n is sufficient to obtain a good approximation.

The non-randomized test $\varphi'(x)$ coincides with the non-randomized procedure which rejects H_0 whenever when the *permutation p-value*

$$\text{ppv}(x) = \sum_\pi I\left\{m(t, d, a_\pi) \geq m(t, d, a)\right\}/n!,$$

is less or equal than α (Lehmann and Romano, 2006, Section 15.2.1). Although $n!$ will typically be too large to compute the $\text{ppv}(x)$ exactly, the benchmark test can be implemented by a *conditional Monte Carlo* approximation. Accordingly, given data x, a large random sample of permutations $\pi_1, \ldots, \pi_B$ ($B = 10^3$, say) is used to estimate the $\text{ppv}(x)$ as $\widehat{\text{ppv}}(x) = \sum_{i=1}^B I\left\{m(t, d, a_{\pi_i}) \geq m(t, d, a)\right\}/B$. The hypothesis H_0 is then rejected if $\widehat{\text{ppv}}(x) \leq \alpha$ (Pesarin and Salmaso, 2010, Section 1.9.3).

6.6 The piecewise exponential model

To implement our maximum-BEP test, we use a *piecewise exponential model* (Benichou and Gail, 1990). The hazard function $h_a(t; \theta)$ is constant over a fixed partition $\tau_0 = 0 < \tau_1 < \cdots < \tau_k < +\infty = \tau_{k+1}$ of the time axis. In particular, $h_a(t; \theta) = \theta_{a,j}$ if $t \in [\tau_{j-1}, \tau_j)$ with $j = 1, \ldots, k+1$, $t \in \mathbb{R}_+$, arms $a = 0, 1$, and $\theta = (\theta_{0,1}, \ldots, \theta_{0,k+1}, \theta_{1,1}, \ldots, \theta_{1,k+1}) \in \Theta = (0, +\infty)^{2(k+1)}$.

The likelihood function of the piecewise exponential model depends on a simple set of sufficient statistics. Given data $x = (t, d, a)$, let $s_{a,j} = \sum_{i=1}^n \max(0, \min(\tau_j - \tau_{j-1}, t_i - \tau_{j-1}))I\{a_i = a\}$ be the total time at risk spent in the interval $[\tau_j, \tau_{j+1})$ by

patients in arm a. Additionally, let $y_{a,j} = \sum_{i=1}^{n} d_i I\{a_i = a,\ \tau_{j-1} \leq t_i < \tau_j\}$ be the number of events observed during $[\tau_{j-1}, \tau_j)$ in arm a. Then, the likelihood is

$$L(\theta; x) = \prod_{a=0}^{1} \prod_{j=1}^{k+1} \theta_{a,j}^{y_{a,j}} \exp(-\theta_{a,j} s_{a,j}).$$

For convenience, we use a conjugate prior $p(\theta)$. This is obtained by letting all $\theta_{a,j}$ be independent and distributed as a gamma random variable with shape parameter $u_{a,j}$ and rate parameter $v_{a,j}$. With this choice, the distribution $p(\theta|x_e)$ presents independent $\theta_{a,j}$ components which are gamma distributed with shape parameter $u_{a,j} + y_{e,a,j}$ and rate parameter $v_{a,j} + s_{e,a,j}$, where the $y_{e,a,j}$ and $s_{e,a,j}$ are the sufficient statistics of x_e. The marginal likelihood $m(x)$ needed to implement the maximum-BEP test can thus be obtained explicitly from Equation 6.6:

$$m(x) \quad = \quad \prod_{a=0}^{1} \prod_{j=1}^{k+1} \left(\frac{v_{a,j} + s_{e,a,j}}{v_{a,j} + s_{e,a,j} + s_{a,j}} \right)^{u_{a,j} + y_{e,a,j} + y_{a,j}} \cdot \frac{\Gamma(u_{a,j} + y_{e,a,j} + y_{a,j})}{\Gamma(u_{a,j} + y_{e,a,j})}, \quad (6.7)$$

where $\Gamma(z)$ is the gamma function.

As an example, Figure 6.1a shows the posterior means of the survival probabilities in the nivolumab or SOC arm of CheckMate 141 obtained from the piecewise exponential model. For all $j = 1, \ldots, k = 4$, we conveniently defined τ_j to be the j-th quintile of the distribution of follow-up times in the SOC arm. In different words, the prior model is chosen by peaking at the early stage trial. Additionally, we specify gamma priors on the $\theta_{a,j}$ with $u_{a,j} = v_{a,j} = 10^{-3}$ for all a and j. The posterior estimates (Figure 6.1a) reflect the delayed separation in the Kaplan-Meier curves, as the estimated survival probabilities diverge only after 4 months of follow-up.

6.7 Application: trials with delayed treatment effects

6.7.1 Simulation study

As an illustration, we use CheckMate 141 data to simulate a large number of phase II and III trials with delayed treatment effects. In these simulations, we compare different tests with respect to their probability of rejecting the hypothesis of no

treatment effects at the end of the phase III trial. We consider Mantel's log-rank test and several others which account for delayed treatment effects: i) a lagged log-rank test that ignores the first 10% of observed follow-up times (Zucker and Lakatos, 1990); ii) the Fleming-Harrington $G^{0,1}$ test, which gives more weight to late events (Fine, 2007); iii) the adaptive log-rank of Yang and Prentice (2010), which weights events according to a preliminary estimate of the hazard functions; and iv) a test of the difference in Restricted Mean Survival Times (RMSTs) across study arms (Huang and Kuan, 2018). We also implement the maximum-BEP test (using the conditional Monte Carlo approach of Section 6.5) based on phase II data. For all tests, we consider $\alpha = 0.05$ and a two-sided alternative hypothesis.

To simulate a trial of size n, we first sample with replacement n patients from the CheckMate 141 data. Then, depending on patient's membership arms, we generate the corresponding survival times from the Kaplan-Meier curves of Figure 6.1a. Assuming a maximum follow-up of 15 months, we generate patient's censoring times by sampling independently from the empirical censoring distribution (Efron, 1981).

Using this approach, we iterate the following steps 10,000 times: i) we simulate a phase II trial of approximately half the size of CheckMate 141 ($n_e = 180$); ii) using the simulated phase II data x_e, we determine the marginal likelihood $m(x)$ for the piecewise exponential model (Equation 6.7); we fix the τ_js at the quintiles of the follow-up times in the SOC arm from x_e and specify the same gamma prior ($u_{a,j} = v_{a,j} = 10^{-3}$) for the parameters $\theta_{a,j}$ as in Section 6.6; iii) we simulate a subsequent phase III study, generating a phase III dataset x with sample size $n = 361$; iv) we apply the test to data x and record the corresponding accept-reject decision. The proportion of rejections across iterations is the Monte Carlo estimate of a test's rejection probability.

Figure 6.1b reports the estimated rejection probabilities for each testing procedure. The maximum-BEP permutation test based on phase II data has the highest probability of rejecting the null hypothesis (approximately 0·90). The $G^{0,1}$ test and the lagged log-rank test have both estimated rejection probabilities of approximately 0·87. The adaptive log-rank and RMST tests have lower rejection probabilities, 0·77 and 0·66 respectively. Mantel's log-rank test has the worst performance, with an estimated rejection probability of 0·60, a third less than the one achieved by our test and the nominal 90% power in the sample size calculations of CheckMate 141 (Ferris et al., 2016). This finding is consistent with previous studies, which highlighted

how the log-rank test may suffer a severe loss of power when treatment effects are delayed (Fine, 2007; Chen, 2013; Alexander et al., 2018).

6.7.2 Robustness analysis

We consider 3 additional simulation scenarios in which the outcome distributions in phase II and III are not identical. In all scenarios, the distribution of the phase II data x_e is the same as in Section 6.7.1, while the distribution of the phase III data x is different. In Scenario 1, the dataset x is generated from the predictive distribution $q(x)$ (see Equation 6.5): a value θ' is first sampled from $p(\theta|x_e)$, then x is generated from the distribution $p_{\theta'}(x)$. Here we assume $r = 2/3$ as in CheckMate 141 and that censoring can only occur after 15 months of follow-up. Proposition 6.5.2 indicates that, in this scenario, our permutation test has the highest expected power. Scenarios 2 and 3 instead represents two settings in which our test may suffer from a loss of power. In Scenario 2, x is generated by a different piecewise exponential model than the one used to construct the benchmark test. The phase III delay in treatment effects is shorter than expected from phase II data. Specifically, x is generated by a model with only one cut-point, fixed at $\tau_1 = 2$ months, whose parameters are set equal to the maximum likelihood estimates obtain from CheckMate 141 data. Scenario 3 is similar, but the cut-point is fixed at $\tau_1 = 8$ months to represent longer phase III delays than those expected from phase II data.

Figure 6.1c shows the results of the robustness analysis. As expected, in Scenario 1 our permutation test has a much higher rejection probability than all other tests (0·98). Instead, its performance is sub-optimal in Scenario 2 and 3. Although in Scenario 3 our permutation test may be considered comparable with the others (rejection probability equal to 0·29), in Scenario 2 it has the lowest rejection probability (0·84, compared to 0·90 for the Mantel's log-rank). These findings support the intuition that the power of the maximum-BEP test depends on how well it is possible to predict the phase III data on the basis of prior information. If the phase II and III trial populations are markedly different, then a test specified using phase II data may perform poorly in the phase III study.

6.8 Generalization to stratified designs

Treatment effects are often expected to vary across patients' groups defined, for example, by gender or biomarkers. In such cases one can stratify patients with respect to covariates measured before randomization. We focus on the primary goal of testing whether the experimental treatment has no effects across all strata or if it is effective at least in some of the strata (alternative hypothesis), for example in one or multiple subgroups defined by a relevant biomarker (Freidlin et al., 2010).

The approach that we discussed can be easily generalized to this setting. For simplicity, we consider the case where each patient $i = 1, \ldots, n$ is categorized by a binary covariate $z_i = 0, 1$, presence ($z_i = 1$) or absence ($z_i = 0$) of a specific marker. Data x thus becomes $x = (t, d, a, z)$, where $z = (z_1, \ldots, z_n) \in \{0, 1\}^n$. Similar to the previous paragraphs we assume that censoring is non-informative and independent of treatment assignments conditionally on $z_1, \ldots, z_n$ (Heitjan and Rubin, 1991).

To illustrate, we specify a piecewise-exponential model $h_a(t; \theta, z)$ for the hazard function in arm $a = 0, 1$ for patients with marker level $z = 0, 1$ (Freidlin et al., 2010): $h_a(t; \theta, z) = \theta_{a,z,j}$ for all $t \in [\tau_{j-1}, \tau_j)$. The prior remains nearly identical to the previous sections. In particular, the marginal likelihood $m(t, d, a, z)$, similar to Equation 6.6, has a closed form expression.

We specify the null hypothesis $H_0 : P \in \mathcal{P}_0'$, where $\mathcal{P}_0'$ is the class of all distributions which are invariant with respect to permutations of the treatment assignment a within the two z groups. More precisely, $P \in \mathcal{P}_0'$ if and only if p, the density of P, satisfy $p(t, d, a, z) = p(t, d, a_\sigma, z)$ for all permutations σ of $(1, \ldots, n)$ such that $z_{\sigma(i)} = z_i$ for all $i = 1, \ldots, n$.

With a simple modification, Proposition 6.5.3 still holds with this new definition of the null hypothesis. Previously, the maximum-BEP permutation test computed the distribution of $m(x)$ under H_0 by considering all permutations of the treatment arm indicators $a_1, \ldots, a_n$. In the stratified case, only permutations σ of $(1, \ldots, n)$ such that $z_{\sigma(i)} = z_i$ for all $i = 1, \ldots, n$ are considered. If $\Sigma(z)$ is the set of all such σ, then the permutation p-value associated to the maximum-BEP test is given by $\text{ppv}(x) = \sum_{\sigma \in \Sigma(z)} I\{m(t, d, a_\sigma, z) \geq m(t, d, a, z)\}/|\Sigma(x)|$, where $|\Sigma(z)| = (\sum_{i=1}^n z_i)!(n - \sum_{i=1}^n z_i)!$.

To provide an example, we simulate 10,000 phase II ($n_e = 180$) and phase III ($n = 361$) trials from CheckMate 141 data in a similar way as in Section 6.7.1.

Differently than in Section 6.7.1, in every trial 50% of patients express ($z_i = 1$) a biomarker predictive of treatment effects (Patel and Kurzrock, 2015). The survival time of a patient in arm $a = 0$ or with marker $z = 0$ is generated from the SOC Kaplan-Meier curve in Figure 6.1a. Instead, the survival time of a patient in arm $a = 1$ with marker $z = 1$ is generated from the nivolumab Kaplan-Meier curve in Figure 6.1a. Censoring times are generated as in Section 6.7.1

In each simulated phase III trial, we tested H_0 in three ways: i) we carried out a test based on the stratified Cox proportional hazards model (a common approach in this setting; c.f. Mehrotra et al., 2012); ii) we performed separate log-rank test in the two marker strata and combined the results using the Bonferroni correction (another common approach; c.f. Freidlin et al., 2014); iii) we implemented our maximum-BEP test using the simulated phase II data. Respectively, the estimated rejection probabilities are 0·18 for the stratified Cox model, 0·26 for the Bonferroni-based test, and 0·49 for our permutation test. These results confirm a substantial benefit in the use of prior data to optimize hypothesis testing.

6.9 Discussion

Data from previous studies should be routinely used to design of late-stage clinical trials. This is especially relevant when standard assumptions, such as the proportional hazards assumption, might not hold. Our approach allows to specify a test for final analyses that accounts for the deviations from proportional hazards suggested by prior data and satisfies the requirements of regulatory agencies (Ventz and Trippa, 2015). The test maximizes a decision-theoretic criteria leveraging on prior data and it is of α-level for an interpretable null hypothesis.

To implement our permutation test, it is necessary to compute the marginal likelihood of the late-stage data. This may be complicated for non-conjugate models. However, many computational methods are available to approximate it (Friel and Wyse, 2012; Pajor et al., 2017).

Although we derived our test assuming a single early-stage dataset, the use of multiple sources of prior data may provide better outcome predictions for late-stage trials. Our approach can incorporate multiple prior datasets using power priors (Ibrahim et al., 2000) or hierarchical models (Spiegelhalter et al., 2004).

Our simulations, based on data from the CheckMate 141 trial, confirm that

weighted log-rank tests can outperform other tests in presence of delayed treatment effects. However, these tests depend on a set of tuning parameters, such as the duration of the lag time for lagged log-rank tests or the ρ and δ coefficients of the $G^{\rho,\delta}$ Fleming-Harrington family, which may be hard to tune. Instead, our approach directly translates early-stage data into a test procedure for the late-stage trial.

Robustness analyses highlight how the performance of our approach is dependent on the consistency of outcome data and the similarity of enrolled populations between phase II and phase III trials. Ensuring the transportability of results to subsequent trials remains a major concern in the design of exploratory clinical trials (Wang et al., 2006).

6.10 Appendix

Denote with $\Pi(x)$ the set of all $\binom{n}{\sum_{i=1}^{n} a_i}$ distinct datasets obtained from $x = (t, d, a)$ by permuting the elements of a in all possible ways. Here, we will assume that when $q(x) > 0$ the inequality $m(x_1) \neq m(x_2)$ holds for all $x_1, x_2 \in \Pi(x)$ such that $x_1 \neq x_2$.

Proposition 6.10.1. Let $\varphi(x)$ be the α-level permutation test of Proposition 6.5.3 and $\varphi'(x)$ its non-randomized version. Then

$$0 \leq BEP_{\varphi} - BEP_{\varphi'} \leq f(\alpha, r, n) = \frac{(1-r)^n}{\alpha} \sum_{s=0}^{n} \left(\frac{r}{1-r} \right)^s$$

Proof. By Proposition 6.5.2, $0 \leq BEP_{\varphi} - BEP_{\varphi'} = E_Q[\varphi(x) - \varphi'(x)] \leq Q(E)$, where E is the set of all x such that $m(x) = m^{(k_\alpha)}(x)$. Proceeding as in Section 5.9 of Lehmann and Romano (2006),

$$Q(E) = \int \frac{\sum_{\sigma} I\left\{m(t, d, a_\sigma) = m^{(k_\alpha)}(x)\right\} q(t, d, a_\sigma)}{\sum_{\sigma} q(t, d, a_\sigma)} dQ(x), \tag{6.8}$$

where both sums extend over all $n!$ permutations σ of $(1, \ldots, n)$. If $q(x) > 0$, then

$$\begin{aligned}
\frac{\sum_\sigma I\left\{m(t,d,a_\sigma) = m^{(k_\alpha)}(x)\right\} q(t,d,a_\sigma)}{\sum_\sigma q(t,d,a_\sigma)} &= \\
&= \frac{\sum_\sigma I\left\{m(t,d,a_\sigma) = m^{(k_\alpha)}(x)\right\} m(t,d,a_\sigma)}{\sum_\sigma m(t,d,a_\sigma)} \\
&\leq \frac{\sum_\sigma I\left\{m(t,d,a_\sigma) = m^{(k_\alpha)}(x)\right\} m(t,d,a_\sigma)}{\sum_\sigma I\left\{m(t,d,a_\sigma) \geq m^{(k_\alpha)}(x)\right\} m(t,d,a_\sigma)} \\
&\leq \frac{\sum_\sigma I\left\{m(t,d,a_\sigma) = m^{(k_\alpha)}(x)\right\} m^{(k_\alpha)}(x)}{\sum_\sigma I\left\{m(t,d,a_\sigma) \geq m^{(k_\alpha)}(x)\right\} m^{(k_\alpha)}(x)} \\
&= \frac{\sum_\sigma I\left\{m(t,d,a_\sigma) = m^{(k_\alpha)}(x)\right\}}{\sum_\sigma I\left\{m(t,d,a_\sigma) \geq m^{(k_\alpha)}(x)\right\}} \\
&= \frac{\#\{j : m^{(j)}(x) = m^{(k_\alpha)}(x)\}}{\#\{j : m^{(j)}(x) \geq m^{(k_\alpha)}(x)\}},
\end{aligned}$$

where the first equality follow because the ratio $q(x)/m(x)$ is invariant with respect to permutations σ of a. Now, by the definitions of $m^{(k_\alpha)}(x)$ and k_α, the denumerator of the last fraction is greater or equal than $\alpha n!$. Instead, the numerator is equal to $n!/\binom{n}{\sum_{i=1}^n a_i}$, as i) $m^{(k_\alpha)}(x) = m(x_\alpha)$ for some $x_\alpha \in \Pi(x)$, ii) for each $x' = (t, d, a') \in \Pi(x)$ there are exactly $n!/\binom{n}{\sum_{i=1}^n a_i}$ permutations σ such that $x' = (t, d, a_\sigma)$, and iii) $m(x)$ assumes distinct values on distinct points of $\Pi(x)$, by assumption. Thus, by Equation 6.8,

$$Q(E) \leq E_Q\left[\frac{1}{\binom{n}{\sum_{i=1}^n a_i}\alpha}\right] = \sum_{s=0}^{n} \frac{1}{\binom{n}{s}\alpha}\binom{n}{s} r^s (1-r)^{n-s} = f(\alpha, r, n).$$

This concludes the proof. □

In general, the approximation provided by Proposition 6.10.1 will be fairly accurate. For example, if $\alpha = 0.05$ and $r = 1/2$, the difference $f(\alpha, r, n)$ between BEP_φ and $BEP_{\varphi'}$ is $2^{-n}(n+1)/\alpha < 10^{-3}$ for all $n \geq 15$. For $r \neq 1/2$ it is $f(\alpha, r, n) = [(1-r)^{n+1} - r^{n+1}]/\alpha(1-2r)$; for $r = 2/3$, the value considered in Section 6.7, $f(\alpha, r, n) < 10^{-3}$ for all $n \geq 25$.

6.11 Available code

I developed R functions to implement the maximum-BEP test of Section 6.5 based on the piecewise-exponential model of Section 6.6. The functions are available,

together with code to reproduce the results of Section 6.7, at `https://github.com/andreaarfe/Bayesian-optimal-tests-non-PH`.

Chapter 7

Concluding remarks

Both the design and data analysis phase are fundamental for the success of biomedical research studies (Fisher et al., 1997; Cox and Donnelly, 2011). In my doctoral research, I aimed to contribute to both phases. I developed novel Bayesian methods for the design and analysis of complex follow-up studies. My interest in the Bayesian paradigm stems for its great usefulness in this setting (Berry and Stangl, 1996; Spiegelhalter et al., 2004; Johnson and de Carvalho, 2015).

During my PhD I developed novel Bayesian approaches to i) model competing risks data, iii) analyze multi-state survival processes, and ii) design of clinical trials with a survival endpoint. The Bayesian perspective was fundamental in each of this. The Bayesian non-parametric paradigm provided me a fresh perspective on modeling complex data without the need for restrictive parametric assumptions (Hjort et al., 2010; Phadia, 2013; Ghosal and van der Vaart, 2017). In addition, the Bayesian paradigm is a natural and effective approach to incorporate data from prior studies in the design of new experiments (Spiegelhalter et al., 2004; Ventz and Trippa, 2015).

Like any Bayesian approach, the methods I developed in this thesis require the specification of prior distributions for model parameters (e.g. the hyper-parameters of the competing-risks regression model in Chapter 4). Ideally, their specification should be based on past data (e.g. as done in Section 4.7.2 for the Weibull regression coefficients), subject-matter knowledge, or expert opinion (Garthwaite et al., 2005; Schmidli et al., 2014; Ibrahim et al., 2015; Kessler et al., 2015). In practice, however, sometimes it may be useful to consider "reference" or "weakly informative" choices (Kass and Wasserman, 1996; Evans and Jang, 2011; Gelman et al., 2013).

In practice, it is often necessary to evaluate the sensitivity of the obtained results with respect to the considered prior distributions, which may be hard to specify precisely Berger, 2013, Section 4.7. Arbitrary or inaccurately specified priors may have an unduly influence on the results (Senn, 2007). The sensitivity of inferences can be assessed by repeating inferences using different prior distributions (Berger, 2013, Section 4.7). To this aim, Besag et al. (1995) detailed an importance-sampling-based approach for assessing prior sensitivity.

I now provide a brief description of future work related to the papers presented in this thesis. In Sections 4.8 and 5.9 above I have already highlighted some future developments related to the papers in Chapters 4 and 5, so I consider them here again only briefly.

With respect to Chapters 4 and 5, I am planning further simulation studies to better appreciate the behavior of the proposed models in comparison with available alternatives under different setups. Several regression models for competing risks data are available in the literature (Crowder, 2012), while Barbu and Limnios (2009) describe a frequentist approach for inference with discrete-time semi-Markov data. In the simulations, methods will be compared with respect to different loss functions (e.g. the classical square error loss; c.f. Berger, 2013) for the problem of estimating relevant model quantities (e.g. the subdistribution function for competing risks, the transition matrix for semi-Markov data).

In a similar vein, I am extending the simulation study of the third paper (Chapter 6) to understand how the piecewise exponential model behaves in different context other than the one characterized delayed treatment effects. Simulations will compare the power of the proposed testing procedure under common patterns of non-proportional hazards (e.g. crossing survival curves; Logan et al., 2008).

I am also generalizing the approach of Chapter 6 using Bayesian non-parametric models in place of the piecewise exponential model. Conceptually, this is possible because the results of Chapter 6 do not require a trial's data-generating distribution to be characterized by a finite-dimensional vector of parameters (c.f. Section 6.3). The proposed approach hinges only the ability to compute the predictive distribution of the new data x conditional on the past data x_e and the truth of the alternative hypothesis of interest. For example, this may be feasible (at least numerically) for models based on the conjugate beta-Stacy process (c.f. 2.3) or Dirichlet Mixture Processes (Antoniak, 1974; Kottas et al., 2005; Ishwaran and James, 2001). For

the former, the predictive distributions can potentially be computed exactly (c.f. Proposition 2.3.1 and the generalization in Section 4.6.1 for discrete competing risks data). For the latter, the predictive distribution can be computed using an approach similar to that of Maceachern and Muller (1998).

I addition, I am extending the procedure of Chapter 6 to other hypothesis testing problems (including tests for non-survival outcomes). Motivated by the problem of testing for treatment effects in a randomized clinical trial, there I considered a non-parametric null hypothesis H_0 defined by a condition of invariance (c.f. Section 6.3). For this specific H_0, it was possible to identify a statistical test with maximum Bayesian expected power using the results of Lehmann and Stein (1949). Results analogous of those in Chapter 6 could potentially be obtained for some other hypotheses - e.g. based on a specific model $p_\theta(x)$, with test for $H_0 : \theta \in \Theta_0$ - but may require different proof techniques. I am currently developing analogous results for single-sample and multi-sample testing problems with general (survival or non-survival) outcome variables.

I will also generalize the approach from Chapter 6 to the design of group-sequential clinical trials based on alternative utility functions (Lewis and Berry, 1994; Bartroff et al., 2012). In this context, data from previous experiments can inform the specification of early stopping rule to accelerate the course of a trial. Evidence of non-proportional hazards from early-stage trials can be used to specify stopping rules for futility or efficacy in confirmatory trials with a survival end-point, a relevant issue for immuno-oncology trials with delayed treatment effects (Zhang and Pulkstenis, 2016).

Finally, starting from developed code (c.f. Sections 4.10 and 6.11), I plan to develop user-friendly R libraries to implement the proposed competing risks regression model (Chapter 4) and maximum-BEP piecewise exponential test for censored data (Chapter 6). Free open-source code will be available on-line (e.g. at the Comprehensive R Archive Network; `https://cran.r-project.org/web/packages/`) to promote the use of Bayesian methods in applications.

Chapter 8

Appendix: other research

During my doctoral studies, I have been working as a visiting research at the Data Science Department of the Dana-Farber Cancer Institute, Boston, Massachusetts (U.S.A.). There, I contributed to several applied research projects, some of which led to the preparation of **3 manuscripts**, **1 accepted for publication and 2 currently under review** in biomedical journals. I describe them briefly below (I omit their full texts here since these are applied works).

Rahman R, Fell G, Ventz S, Arfè A, Vanderbeek A, Trippa L, Alexander B. Deviation from the proportional hazards assumption in randomised phase 3 oncology clinical trials: prevalence, associations, and implications. (Clinical Cancer Research 2019, forthcoming. On-line pre-print: `https://clincancerres.aacrjournals.org/content/early/2019/07/25/1078-0432.CCR-18-3999` *)*

Deviations from proportional hazards, which may be more prevalent in the era of precision medicine and immunotherapy, can lead to under-powered trials or misleading conclusions. We used a meta-analytic patient-level approach to estimate deviations from proportional hazards across cancer trials, investigate associated factors, and evaluate alternative analytic approaches for future trial designs. From 152 oncology trials, we obtained data on 129,401 patients, which we re-analyzed to evaluate Deviations from proportional hazards and specific alternative statistical procedures. Among the included trials, 75 (24.7%) exhibited evidence of non-proportional hazards. We found non-proportional hazards were more common for

cancer immunotherapy trials and trials with composite endpoints (e.g. progression free survival). For the design and analysis of oncology trials, we provide quantitative justification for the use of statistical methods that do not rely on the proportional hazards assumption.

Arfè A, Fell G, Alexander B, Awad M, Rodig S, Trippa L. Pooled analysis of Programmed Death Factor Ligand 1 expression as a predictive biomarker using individual data on 7,918 randomized study patients. (Submitted manuscript.)

Programmed Cell Death Factor Ligand 1 (PD-L1) expression is one of the most studied biomarkers to predict the efficacy of immune checkpoint inhibitors (ICIs), but its clinical significance is controversial. In this study, we reconstructed, pooled, and analyzed individual-level data on 7,918 cancer patients from 14 randomized clinical trials. We estimated i) the distribution of PD-L1 expression scores (i.e. tumor proportion score or combined proportion score), and ii) the relationship between PD-L1 levels and ICIs' impact on overall survival (OS). ICIs' effects were quantified using differences in 24-months restricted mean survival times, i.e. the increase in 2-years life expectancy associated with ICI therapy. In a simulation study, we show how estimates of the distribution of PD-L1 scores and PD-L1-specific treatment effects like ours can be used to improve future trials designs to detect ICIs' benefits. Our findings suggest that the practice of dichotomizing the range of PD-L1 expression scores is inadequate for patient stratification.

Spring LM, Fell G*, Arfè A*, Sharma C, Greenup R, Reynolds KL, Smith BL, Alexander B, Moy B, Isakoff SJ, Parmigiani G, Trippa L, Bardia A. Pathological complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. *Co-primary author. (Submitted manuscript.)*

This paper deals with the use of pathological complete response (pCR) as a surrogate outcome to accelerate the clinical evaluation of breast cancer therapies in the neo-adjuvant setting. While the prognostic significance of pathological complete response (pCR) after neoadjuvant chemotherapy is relatively well established, the impact of adjuvant therapy in modulating the relationship between pCR and long

term outcomes is less clear. To assess the association between pCR and survival, we extracted patient-level data for over 20,000 breast cancer patients from 52 randomized clinical trials of novel neoadjuvant therapies. We quantified the association between pCR and survival using Bayesian hierarchical models for censored data, including pCR as a predictor. Our results suggest that achieving pCR following neoadjuvant chemotherapy is associated with significantly improved survival, particularly for triple negative and HER2+ breast cancer. Our results also suggest that adjuvant chemotherapy could potentially be abbreviated in certain circumstances.

Bibliography

Aalen, O., Borgan, O. and Gjessing, H. (2008) *Survival and event history analysis: a process point of view*. New York: Springer Science & Business Media.

Agresti, A. (2003) *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Alexander, B. M., Schoenfeld, J. D. and Trippa, L. (2018) Hazards of hazard ratios—deviations from model assumptions in immunotherapy. *New England Journal of Medicine*, **378**, 1158–1159.

Allison, P. D. (1982) Discrete-time methods for the analysis of event histories. *Sociological methodology*, **13**, 61–98.

Amerio, E., Muliere, P. and Secchi, P. (2004) Reinforced urn processes for modeling credit default distributions. *International Journal of Theoretical and Applied Finance*, **7**, 407–423.

Andersen, P. K., Abildstrom, S. Z. and Rosthøj, S. (2002) Competing risks as a multi-state model. *Statistical Methods in Medical Research*, **11**, 203–215.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (2012) *Statistical models based on counting processes*. New York: Springer Science & Business Media.

Andersen, P. K. and Keiding, N. (2002) Multi-state models for event history analysis. *Statistical methods in medical research*, **11**, 91–115.

Anscombe, F. (1963) Sequential medical trials. *Journal of the American Statistical Association*, **58**, 365–383.

Antoniak, C. E. (1974) Mixtures of dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, **2**, 1152–1174.

Arfè, A., Peluso, P. and Muliere, P. (2018) The semi-markov beta-stacy process: a bayesian non-parametric prior for semi-markov processes. *Submitted manuscript.* ArXiv manuscript: http://arxiv.org/abs/1812.00260.

Ashby, D. and Smith, A. F. (2000) Evidence-based medicine as bayesian decision-making. *Statistics in medicine*, **19**, 3291–3305.

Bacallado, S., Favaro, S. and Trippa, L. (2013) Bayesian nonparametric analysis of reversible Markov chains. *The Annals of Statistics*, 870–896.

Barbu, V., Boussemart, M. and Limnios, N. (2004) Discrete-time semi-markov model for reliability and survival analysis. *Communications in Statistics-Theory and Methods*, **33**, 2833–2868.

Barbu, V. and Limnios, N. (2009) *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis*. Lecture Notes in Statistics. Springer New York.

Barthel, F.-S., Babiker, A., Royston, P. and Parmar, M. (2006) Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statistics in medicine*, **25**, 2521–2542.

Bartroff, J., Lai, T. and Shih, M. (2012) *Sequential Experimentation in Clinical Trials: Design and Analysis*. Springer Series in Statistics. Springer New York.

Benichou, J. and Gail, M. H. (1990) Estimates of absolute cause-specific risk in cohort studies. *Biometrics*, 813–826.

Berger, J. (2013) *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York.

Berger, M. and Schmid, M. (2017) Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 1471082X17748084.

Bernardo, J. M. and Smith, A. F. (2000) *Bayesian theory*. John Wiley & Sons.

Berry, D. and Stangl, D. (1996) *Bayesian Biostatistics*. Statistics: A Series of Textbooks and Monographs. CRC Press.

Berry, D. A. and Ho, C.-H. (1988) One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics*, 219–227.

Berry, S. M., Carlin, B. P., Lee, J. J. and Muller, P. (2010) *Bayesian adaptive methods for clinical trials*. CRC press.

Berry, S. M. and Kadane, J. B. (1997) Optimal bayesian randomization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 813–819.

Besag, J., Green, P., Higdon, D., Mengersen, K. et al. (1995) Bayesian computation and stochastic systems. *Statistical science*, **10**, 3–41.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via polya urn schemes. *Ann. Statist.*, **1**, 353–355.

Brody, T. (2016) *Clinical Trials: Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines*. Elsevier Science.

Brown, B. W., Herson, J., Atkinson, E. N. and Rozell, M. E. (1987) Projection from previous studies: a bayesian and frequentist compromise. *Controlled clinical trials*, **8**, 29–44.

Bulla, J. and Bulla, I. (2006) Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics & Data Analysis*, **51**, 2192–2209.

Bulla, P. and Muliere, P. (2007) Bayesian nonparametric estimation for reinforced markov renewal processes. *Statistical Inference for Stochastic Processes*, **10**, 283–303.

Bulla, P., Muliere, P. and Walker, S. (2009) A Bayesian nonparametric estimator of a multivariate survival function. *Journal of Statistical Planning and Inference*, **139**, 3639–3648.

Carlin, B. P., Kadane, J. B. and Gelfand, A. E. (1998) Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, 964–975.

Caron, F., Neiswanger, W., Wood, F., Doucet, A. and Davy, M. (2017) Generalized pólya urn for time-varying pitman-yor processes. *Journal of Machine Learning Research*, **18**, 1–32.

Çinlar, E. (2011) *Probability and Stochastics*. New York: Springer.

Cellamare, M., Milstein, M., Ventz, S., Baudin, E., Trippa, L. and Mitnick, C. (2016) Bayesian adaptive randomization in a clinical trial to identify new regimens for mdr-tb: the endtb trial. *The International Journal of Tuberculosis and Lung Disease*, **20**, S8–S12.

Chae, M., Weißbach, R., Cho, K. H. and Kim, Y. (2013) A mixture of beta–dirichlet processes prior for Bayesian analysis of event history data. *Journal of the Korean Statistical Society*, **42**, 313–321.

Chen, L.-A., Hung, H.-N. and Chen, C.-R. (2007) Maximum average-power (map) tests. *Communications in Statistics—Theory and Methods*, **36**, 2237–2249.

Chen, T.-T. (2013) Statistical issues and challenges in immuno-oncology. *Journal for immunotherapy of cancer*, **1**, 18.

Chevret, S. (2012) Bayesian adaptive clinical trials: a dream for statisticians only? *Statistics in medicine*, **31**, 1002–1013.

Chow, S., Wang, H. and Shao, J. (2007) *Sample Size Calculations in Clinical Research, Second Edition*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis.

Cifarelli, D. and Regazzini, E. (1982) Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability. *Exchangeability in Probability and Statistics*, 185–205.

Cifarelli, D. M. and Regazzini, E. (1978) Problemi statistici non parametrici in condizioni di scambiabilità parziale: impiego di medie associative. *Tech. rep.*, Quaderni Istituto di Matematica Finanziaria dell'Università di Torino. English translation, "Nonparametric statistical problems under partial exchangeability. The role of associative means", available at . http://www-dimat.unipv.it/eugenioconference/eugenio.html (last accessed: 24 March 2018).

— (1996) De Finetti's contribution to probability and statistics. *Statistical Science*, **11**, 253–282.

Çinlar, E. (1969) Markov renewal theory. *Advances in Applied Probability*, **1**, 123–187.

Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, **64**, 194–206.

Coppersmith, D. and Diaconis, P. (1986) Random walk with reinforcement. unpublished manuscript.

Cornfield, J. (1966) Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, **20**, 18–23.

Cox, D. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 87–22.

Cox, D. and Donnelly, C. (2011) *Principles of Applied Statistics*. Cambridge University Press.

Crowder, M. J. (2012) *Multivariate survival analysis and competing risks*. Boca Raton, Florida: CRC Press.

Damien, P., Laud, P. W. and Smith, A. F. (1995) Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 547–563.

Dawid, A. P. (1988) Symmetry models and hypotheses for structured data layouts. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–34.

De Blasi, P. and Hjort, N. L. (2007) Bayesian survival analysis in proportional hazard models with logistic relative risk. *Scandinavian Journal of Statistics*, **34**, 229–257.

de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, **7**, 1–68. English translation, "Foresight: its Logical

Laws, Its Subjective Sources", in H. E. Kyburg and H. E. Smokler (editors), Studies in Subjective Probability. New York: Wiley, 1964.

DeGroot, M. H. (2005) *Optimal statistical decisions*, vol. 82. John Wiley & Sons.

Diaconis, P. (1988) Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics 3* (eds. J. Bernardo, M. DeGroot, D. Lindley and A. Smith), vol. 3, 111–125. Oxford University Press Oxford,, UK.

Diaconis, P. and Freedman, D. (1980) De Finetti's theorem for Markov chains. *The Annals of Probability*, **8**, 115–130.

Ding, M., Rosner, G. L. and Müller, P. (2008) Bayesian optimal design for phase ii screening trials. *Biometrics*, **64**, 886–894.

Doksum, K. (1974) Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 183–201.

Drzewiecki, K., Ladefoged, C. and Christensen, H. (1980) Biopsy and prognosis for cutaneous malignant melanomas in clinical stage I. *Scandinavian Journal of Plastic and Reconstructive Surgery*, **14**, 141–144.

Dunson, D. B. (2001) Commentary: practical advantages of bayesian analysis of epidemiologic data. *American journal of Epidemiology*, **153**, 1222–1226.

Efron, B. (1981) Censored data and the bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

Epifani, I., Fortini, S. and Ladelli, L. (2002) A characterization for mixtures of semi-Markov processes. *Statistics & Probability Letters*, **60**, 445–457.

Epifani, I., Lijoi, A. and Pruenster, I. (2003) Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, **90**, 791–808.

Evans, M. and Jang, G. H. (2011) Weak informativity and the information in one prior relative to another. *Statistical Science*, **26**, 423–439.

Ferguson, T. (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615–629.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Ferguson, T. S. and Phadia, E. G. (1979) Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 163–186.

Ferris, R. L., Blumenschein Jr, G., Fayette, J., Guigay, J., Colevas, A. D., Licitra, L., Harrington, K., Kasper, S., Vokes, E. E., Even, C. et al. (2016) Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine*, **375**, 1856–1867.

Fine, G. D. (2007) Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug information journal*, **41**, 535–539.

Fine, J. P. (1999) Analysing competing risks data with transformation models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, **61**, 817–830.

Fine, J. P. and Gray, R. J. (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 496–509.

Fisher, L., Velle, G. v. and Elsayed, E. (1997) Biostatistics: a methodology for the health sciences. *IIE Transactions*, **29**, 799.

Fisher, R. A. (1935) *The design of experiments.* Oliver & Boyd.

Fleming, T. and Harrington, D. (2011) *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics. Wiley.

Fortini, S. and Petrone, S. (2012) Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, **26**, 423–449.

Fortini, S., Petrone, S. and Sporysheva, P. (2016) On a notion of partially conditionally identically distributed sequences. [arXiv:math.PA/1608.00471v1].

Freidlin, B., Korn, E. L. and Gray, R. (2014) Marker sequential test (mast) design. *Clinical trials*, **11**, 19–27.

Freidlin, B., McShane, L. M. and Korn, E. L. (2010) Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, **102**, 152–160.

Friel, N. and Wyse, J. (2012) Estimating the evidence–a review. *Statistica Neerlandica*, **66**, 288–308.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005) Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–701.

Ge, M. and Chen, M.-H. (2012) Bayesian inference of the fully specified subdistribution model for survival data with competing risks. *Lifetime Data Analysis*, **18**, 339–363.

Geisser, S. (1993) *Predictive inference*, vol. 55. CRC press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013) *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: Taylor & Francis, third edition edn.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calcualting posterior moments. In *Bayesian Statistics 4* (eds. J. Bernardo, J. Berger, A. Dawid and S. A.F.M.), 169–193. Oxford, UK: Clarendon Press.

Ghosal, S. and van der Vaart, A. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gill, R. D., Johansen, S. et al. (1990) A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, **18**, 1501–1555.

Gómez, G., Gómez-Mateu, M. and Dafni, U. (2014) Informed choice of composite end points in cardiovascular trials. *Circulation: Cardiovascular Quality and Outcomes*, **7**, 170–178.

Good, P. (2006) *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer New York.

Guo, S. and Lin, D. (1994) Regression analysis of multivariate grouped survival data. *Biometrics*, 632–639.

Guyot, P., Ades, A., Ouwens, M. J. and Welton, N. J. (2012) Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, **12**, 9.

Heitjan, D. F. (1993) Ignorability and coarse data: Some biomedical examples. *Biometrics*, **49**, 1099–1109.

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *The Annals of Statistics*, **19**, 2244–2253.

Hinchliffe, S. R. and Lambert, P. C. (2013) Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology*, **13**, 13.

Hjort, N. L. (1990) Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, **18**, 1259–1294.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010) *Bayesian nonparametrics*. Cambridge, UK: Cambridge University Press.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011) Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, **67**, 1047–1056.

van Houwelingen, H. C., van de Velde, C. J. and Stijnen, T. (2005) Interim analysis on survival data: its potential bias and how to repair it. *Statistics in medicine*, **24**, 2823–2835.

Huang, B. and Kuan, P.-F. (2018) Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharmaceutical statistics*, **17**, 202–213.

Huber, C., Pons, O. and Heutte, N. (2004) Independent competing risks versus a general semi-Markov model. Application to heart transplant data. *Lifetime Data Analysis*, 1.

Ibrahim, J. G., Chen, M.-H., Gwon, Y. and Chen, F. (2015) The power prior: theory and applications. *Statistics in medicine*, **34**, 3724–3749.

Ibrahim, J. G., Chen, M.-H. et al. (2000) Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

James, L. F. (2002) Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*.

— (2006) Poisson calculus for spatial neutral to the right processes. *The annals of Statistics*, **34**, 416–440.

Janssen, J. and Manca, R. (2007) *Semi-Markov Risk Models for Finance, Insurance and Reliability*. Springer US.

Jeong, J.-H. and Fine, J. P. (2007) Parametric regression on cumulative incidence function. *Biostatistics*, **8**, 184–196.

Johnson, W. O. and de Carvalho, M. (2015) Bayesian nonparametric biostatistics. In *Nonparametric Bayesian Inference in Biostatistics*, 15–54. Springer.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The statistical analysis of failure time data*. Hoboken, New Jersey: John Wiley & Sons, 2nd edition edn.

Kallenberg, O. (2006) *Foundations of Modern Probability*. Probability and Its Applications. Springer New York.

— (2010) *Probabilistic Symmetries and Invariance Principles*. Probability and Its Applications. Springer New York.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Kessler, D. C., Hoff, P. D. and Dunson, D. B. (2015) Marginally specified priors for non-parametric bayesian estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**, 35–58.

Kim, H. T. and Gray, R. (2012) Three-component cure rate model for nonproportional hazards alternative in the design of randomized clinical trials. *Clinical Trials*, **9**, 155–163.

Kingman, J. (1967) Completely random measures. *Pacific Journal of Mathematics*, **21**, 59–78.

Ko, S. I., Chong, T. T., Ghosh, P. et al. (2015) Dirichlet process hidden Markov multiple change-point model. *Bayesian Analysis*, **10**, 275–296.

Kottas, A., Müller, P. and Quintana, F. (2005) Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, **14**, 610–625.

Larson, M. G. and Dinse, G. E. (1985) A mixture model for the regression analysis of competing risks data. *Applied Statistics*, **34**, 201–211.

Lau, B., Cole, S. R. and Gange, S. J. (2009) Competing risk regression models for epidemiologic data. *American journal of epidemiology*, **2**, 244–256.

Lawless, J. F. (2011) *Statistical models and methods for lifetime data*. Hoboken, New Jersey: John Wiley & Sons.

Lee, J. (2007) Sampling methods of neutral to the right processes. *Journal of Computational and Graphical Statistics*, **16**, 656–671.

Lee, K. M. and Wason, J. (2018) Design of experiments for a confirmatory trial of precision medicine. *Journal of Statistical Planning and Inference*.

Lehmann, E. and Romano, J. (2006) *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer New York.

Lehmann, E. L. and Stein, C. (1949) On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, **20**, 28–45.

Lewis, R. J. and Berry, D. A. (1994) Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*, **89**, 1528–1534.

Lindley, D. (1994) Discussion of "bayesian approaches to randomized trials" by d.j. spiegelhalter, l.s. freedman, and m.k.b. parmar. *Journal of the Royal Statistical Society. Series A (Statistics in Society).*, **157**, 393.

Lindley, D. and Smith, A. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 1–41.

Lindley, D. V. (1997) The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **46**, 129–138.

Liu, F. (2018) Assessment of bayesian expected power via bayesian bootstrap. *Statistics in medicine*, **37**, 3471–3485.

Logan, B. R., P. Klein, J. and Zhang, M. (2008) Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*, **64**, 733–740.

Maceachern, S. N. and Muller, P. (1998) Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

Mahmoud, H. (2008) *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Martin, A. D., Quinn, K. M. and Park, J. H. (2011) MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, **42**, 22.

Masala, G. (2013) Hurricane lifespan modeling through a semi-Markov parametric approach. *Journal of Forecasting*, **32**, 369–384.

Mauldin, R. D., Sudderth, W. D. and Williams, S. (1992) Polya trees and random distributions. *The Annals of Statistics*, **20**, 1203–1221.

Mehrotra, D. V., Su, S.-C. and Li, X. (2012) An efficient alternative to the stratified cox model analysis. *Statistics in medicine*, **31**, 1849–1856.

Mehta, R. S., Barlow, W. E., Albain, K. S., Vandenberg, T. A., Dakhil, S. R., Tirumali, N. R., Lew, D. L., Hayes, D. F., Gralow, J. R., Livingston, R. B. et al. (2012) Combination anastrozole and fulvestrant in metastatic breast cancer. *New England Journal of Medicine*, **367**, 435–444.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suarez, C. and Andersen, P. K. (2009) Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, **18**, 195–222.

Mezzetti, M., Muliere, P. and Bulla, P. (2007) An application of reinforced urn processes to determining maximum tolerated dose. *Statistics & probability letters*, **77**, 740–747.

Mihram, G. A. and Hultquist, R. A. (1967) A bivariate warning-time/failure-time distribution. *Journal of the American Statistical Association*, **62**, 589–599.

Mira, A. and Petrone, S. (1996) Bayesian hierarchical nonparametric inference for change-point problems. In *Bayesian Statistics* (eds. J. M. Bernardo, J. Berger, A. Dawid and A. Smith), no. 5, 693–703. Oxford University Press.

Mitchell, C., Hudgens, M., King, C., Cu-Uvin, S., Lo, Y., Rompalo, A., Sobel, J. and Smith, J. (2011) Discrete-time semi-Markov modeling of human papillomavirus persistence. *Statistics in medicine*, **30**, 2160–2170.

Mitra, R. and Müller, P. (2015) *Nonparametric Bayesian Inference in Biostatistics*. Frontiers in Probability and the Statistical Sciences. Springer International Publishing.

Muliere, P., Paganoni, A. M. and Secchi, P. (2006) A randomly reinforced urn. *Journal of Statistical Planning and Inference*, **136**, 1853–1874.

Muliere, P. and Petrone, S. (1993) A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Statistical Methods & Applications*, **2**, 349–364.

Muliere, P. and Scarsini, M. (1985) Change-point problems: A and Bayesian nonparametric approach. *Aplikace Matematiky*, **30**, 397–402.

Muliere, P., Secchi, P. and Walker, S. (2000) Urn schemes and reinforced random walks. *Stochastic Processes and their Applications*, **88**, 59–78.

— (2005) Partially exchangeable processes indexed by the vertices of a k-tree constructed via reinforcement. *Stochastic processes and their applications*, **115**, 661–677.

Muliere, P., Secchi, P. and Walker, S. G. (2003) Reinforced random processes in continuous time. *Stochastic Processes and their Applications*, **104**, 117–130.

Muliere, P. and Walker, S. (2000) Neutral to the right processes from a predictive perspective: a review and new developments. *Metron*, **58**, 13–30.

Müller, P., Berry, D. A., Grieve, A. P., Smith, M. and Krams, M. (2007) Simulation-based sequential bayesian design. *Journal of statistical planning and inference*, **137**, 3140–3150.

Müller, P. and Mitra, R. (2013) Bayesian nonparametric inference–why and how. *Bayesian Analysis*, **8**, 269–302.

Müller, P., Quintana, F., Jara, A. and Hanson, T. (2015) *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Springer International Publishing.

Nakagawa, T. and Osaki, S. (1975) The discrete weibull distribution. *IEEE Transactions on Reliability*, **24**, 300–301.

Ng, K. W., Tian, G.-L. and Tang, M.-L. (2011) *Dirichlet and related distributions: Theory, methods and applications*. Chichester, England: John Wiley & Sons.

Nieto-Barajas, L. E. and Walker, S. G. (2002) Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics*, **29**, 413–424.

O'Hagan, A., Stevens, J. W. and Campbell, M. J. (2005) Assurance in clinical trial design. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, **4**, 187–201.

Orbanz, P. and Roy, D. M. (2015) Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, **37**, 437–461.

Paganoni, A. M. and Secchi, P. (2004) Interacting reinforced-urn systems. *Advances in applied probability*, **36**, 791–804.

Pajor, A. et al. (2017) Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, **12**, 261–287.

Parmigiani, G. and Inoue, L. (2009) *Decision Theory: Principles and Approaches*. Wiley Series in Probability and Statistics. Wiley.

Patel, S. P. and Kurzrock, R. (2015) Pd-l1 expression as a predictive biomarker in cancer immunotherapy. *Molecular cancer therapeutics*, **14**, 847–856.

Patwardhan, A. S., Kulkarni, R. B. and Tocher, D. (1980) A semi-Markov model for characterizing recurrence of great earthquakes. *Bulletin of the seismological society of America*, **70**, 323–347.

Peluso, S., Chib, S., Mira, A. et al. (2018) Semiparametric multivariate and multiple change-point modeling. *Bayesian Analysis*.

Peluso, S., Mira, A. and Muliere, P. (2015) Reinforced urn processes for credit risk models. *Journal of Econometrics*, **184**, 1–12.

Peluso, S., Mira, A., Muliere, P. et al. (2017) Learning vs earning trade-off with missing or censored observations: The two-armed bayesian nonparametric beta-stacy bandit problem. *Electronic Journal of Statistics*, **11**, 3368–3406.

Pemantle, R. (1988) *Random processes with reinforcement*. Ph.D. thesis, Massachussets Institute of Technology.

— (2007) A survey of random processes with reinforcement. *Probabability Surveys*, **4**, 1–79.

Pesarin, F. and Salmaso, L. (2010) *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley Series in Probability and Statistics. Wiley.

Phadia, E. G. (2013) *Prior processes and their applications*. New York: Springer.

Phelan, M. J. (1990) Bayes estimation from a Markov renewal process. *The Annals of Statistics*, **18**, 603–616.

Phillips, A. and Haudiquet, V. (2003) Ich e9 guideline 'statistical principles for clinical trials': a case study. *Statistics in medicine*, **22**, 1–11.

Pintilie, M. (2006) *Competing risks: a practical perspective*. Chichester, England: John Wiley & Sons.

Pitman, J. (1996) Some developments of the blackwell-macqueen urn scheme. *Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell*, **30**, 245.

Putter, H., Fiocco, M. and Geskus, R. (2007) Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389–2430.

Rigat, F. and Muliere, P. (2012) Nonparametric survival regression using the beta-Stacy process. *Journal of Statistical Planning and Inference*, **142**, 2688–2700.

Robert, C. (2007) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York.

Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York: Springer, 2nd edn.

Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., Lichinitser, M., Dummer, R., Grange, F., Mortier, L. et al. (2015) Improved overall survival in melanoma with combined dabrafenib and trametinib. *New England Journal of Medicine*, **372**, 30–39.

Royston, P. and Parmar, M. K. (2013) Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, **13**, 152.

Ruggiero, M. and Walker, S. G. (2009) Bayesian nonparametric construction of the fleming-viot process with fertility selection. *Statistica Sinica*, 707–720.

Sato, K.-I. (1999) *Lévy processes and infinitely divisible distributions*. Cambridge, UK: Cambridge University Press.

Satten, G. A. and Sternberg, M. R. (1999) Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics*, **55**, 507–513.

Scheike, T. H. and Zhang, M.-J. (2011) Analyzing competing risk data using the r timereg package. *Journal of statistical software*, **38**.

Scheike, T. H., Zhang, M.-J. and Gerds, T. A. (2008) Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, **95**, 205–220.

Schiffman, M., Castle, P. E., Maucort-Boulch, D., Wheeler, C. M., of Undetermined Significance/Low-Grade Squamous Intraepithelial Lesions Triage Study) Group, A. A. S. C. and Plummer, M. (2007) A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion. *The Journal of infectious diseases*, **195**, 1582–1589.

Schmid, M., Küchenhoff, H., Hoerauf, A. and Tutz, G. (2016) A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in medicine*, **35**, 734–751.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014) Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, **70**, 1023–1032.

Senn, S. (2007) Trying to be precise about vagueness. *Statistics in Medicine*, **26**, 1417–1430.

Shen, W., Sakamoto, N. and Yang, L. (2016) Melanoma-specific mortality and competing mortality in patients with non-metastatic malignant melanoma: a population-based analysis. *BMC Cancer*, **16**, 413.

Singpurwalla, N. D. (1988) Foundational issues in reliability and risk analysis. *SIAM Review*, **30**, 264–282.

— (2006) *Reliability and risk: a Bayesian perspective*. Chichester, England: John Wiley & Sons.

Sivazlian, B. (1981) On a multivariate extension of the gamma and beta distributions. *SIAM Journal on Applied Mathematics*, **41**, 205–209.

Smith, A. (1975) A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, **62**, 407–416.

Spiegelhalter, D., Abrams, K. and Myles, J. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. Wiley.

Spiegelhalter, D. and Freedman, L. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in medicine*, **5**, 1–13.

Stallard, N., Thall, P. F. and Whitehead, J. (1999) Decision theoretic designs for phase ii clinical trials with multiple outcomes. *Biometrics*, **55**, 971–977.

Tamm, M. and Hilgers, R.-D. (2014) Chronological bias in randomized clinical trials arising from different types of unobserved time trends. *Methods of information in medicine*, **53**, 501–510.

Thörn, M., Ponté, F., Bergström, R., Sparén, P. and Adami, H.-O. (1994) Clinical and histopathologic predictors of survival in patients with malignant melanoma: a population-based study in Sweden. *JNCI: Journal of the National Cancer Institute*, **86**, 761–769.

Trippa, L., Rosner, G. L. and Müller, P. (2012) Bayesian enrichment strategies for randomized discontinuation trials. *Biometrics*, **68**, 203–211.

Tutz, G. and Schmid, M. (2016) *Modeling Discrete Time-to-Event Data*. Springer Series in Statistics. Springer.

US Food and Drug Administration (1998) Guidance for industry: Statistical principles for clinical trials. Accessed October 5, 2017.

Ventz, S., Cellamare, M., Bacallado, S. and Trippa, L. (2018) Bayesian uncertainty directed trial designs. *Journal of the American Statistical Association*, 1–38.

Ventz, S., Parmigiani, G. and Trippa, L. (2017) Combining bayesian experimental designs and frequentist data analyses: motivations and examples. *Applied Stochastic Models in Business and Industry*, **33**, 302–313.

Ventz, S. and Trippa, L. (2015) Bayesian designs and the control of frequentist characteristics: a practical solution. *Biometrics*, **71**, 218–226.

Wade, S., Mongelluzzo, S., Petrone, S. et al. (2011) An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, **6**, 359–385.

Walker, S. and Damien, P. (1998a) A full Bayesian non-parametric analysis involving a neutral to the right process. *Scandinavian Journal of Statistics*, **25**, 669–680.

— (1998b) Sampling methods for bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, 243–254. Springer.

Walker, S. and Muliere, P. (1997) Beta-Stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics*, **25**, 1762–1780.

— (1999) A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *The Annals of Statistics*, **27**, 589–599.

Wang, S.-J., Hung, H. J. and O'Neill, R. T. (2006) Adapting the sample size planning of a phase iii trial based on phase ii data. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, **5**, 85–97.

Wechsler, S. (1993) Exchangeability and predictivism. *Erkenntnis*, **38**, 343–350.

Wolbers, M., Koller, M. T., Witteman, J. C. and Steyerberg, E. W. (2009) Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*, **20**, 555–561.

Yang, S. and Prentice, R. (2010) Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, **66**, 30–38.

Zabell, S. L. et al. (1982) W.E. Johnson's "sufficientness" postulate. *The annals of statistics*, **10**, 1090–1099.

Zhang, J. and Pulkstenis, E. (2016) Sample size and power of survival trials in group sequential design with delayed treatment effect. *Statistics in Biopharmaceutical Research*, **8**, 268–275.

Zhao, L. and Hu, X. J. (2013) Estimation with right-censored observations under a semi-Markov model. *Canadian Journal of Statistics*, **41**, 237–256.

Zucker, D. M. and Lakatos, E. (1990) Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*, **77**, 853–864.